Multiple Linear Regression Viewpoints

A Publication sponsored by the American Educational Research Association's Special Interest Group on Multiple Linear Regression: The General Linear Model

MLRV

Volume 26 • Number 2 • Fall 2000

Table of Contents

In Memoriam: Max R. Martin **1**
Nancy K. Martin, University of Texas-San Antonio

Outliers Lie: An Illustrative Example of Identifying Outliers and Applying Robust Models **2**
Karl Ho, University of North Texas
Jimmie R. Naugher, University of North Texas

Problems with Probabilistic Hindsight: A Comparison of Methods for Retrospective Statistical Power Analysis **7**
Jeffrey Kromrey, University of South Florida
Kristine Y. Hogarty, University of South Florida

Evaluating Univariate and Bivariate Normality Using Graphical Procedures **15**
Tom Burdenski, Texas A & M University

Using Partial Residual Plots in Assessing and Improving the Construct Validity of Multiple Regression Models **29**
Cam-Loi Huynh, University of Manitoba

Rasch Measurement Instead of Regression **36**
Benjamin D. Wright, University of Chicago
Kyle Perkins, Southern Illinois University
J. Kevin Dorsey, Southern Illinois University

Multiple Regression with WINSTEPS A Rasch Solution to Regression Confusion **42**
Benjamin D. Wright, University of Chicago

In Memoriam: Max R. Martin

Nancy K. Martin, University of Texas at San Antonio

When I was asked to write about my dear friend Max Martin, I was honored to be able to remember him in this way. I am however overwhelmed as I stare at the blinking cursor and try to figure out what to say. I've known Max for years. It seems like this task would be easy, but the idea of a memoriam for Max is surreal. All of us who knew him are in shock over his untimely passing. Also, it's hard to know what to say about someone like Max. How do I sum up such a special life in just a few words? No doubt, an impossible task; still, I will try. Max passed away on August 18, 2000 at age 52 – eight days after suffering a massive heart attack at home. Prior to his death, Max used his amazing computer and statistical skills as the senior evaluator in the Research and Evaluation Systems Technology Department for one of the poorest school districts in Texas. Ironically, Max began his career as a chemical engineer. He could have done anything he wanted to do, been anything he wanted to be, but after a stint as a math teacher at a Catholic school, his career path was clear. Education was his calling and it was that road that crossed with mine.

I first came to know Max when we were enrolled in the same doctoral program at Texas Tech University in the early 1980s. We struggled through graduate school together (I struggled more than he) and we formed that special bond that only fellow doctoral students can understand.

We had the same last name and that often led to confusion. His mail was in my mailbox; his student messages were on my answering machine at home - "Would you please tell your husband . . . ?" The idea that we were husband and wife was a natural assumption and we had a lot of fun with it over the years. By coincidence, we both ended up living in San Antonio and we both have served as past Presidents of the Southwest Educational Research Association, so the confusion (and fun) continued.

There's nothing like a good story to provide a composite picture of someone, especially Max. Max and I had the same dissertation chairman, Dr. Paul Dixon. Dr. Paul Dixon was tough as nails, but Max and I were both wise enough to realize a smart choice when we saw one. Everyone knew two things before enrolling in one of Paul's classes. First, be ready to work hard, and second, we were going to learn more than we probably wanted to know about the subject matter. Paul's take-home final exam in Learning Theory was inspired and infamous. The final exam was a dialogue between learning theorists we studied during the semester with questions in between that we had to respond to by Wertheimer, Pavlov, Skinner, and others. The exam was extremely difficult and we were required to work independently. We all scurried home to work into the wee hours of the morning drinking gallons of coffee for days on end. Finally, the exam due date arrived. Max's answers were brilliantly composed, thoughtful and insightful, but he had an added surprise. His answers were not in English, instead with a little help from his friends, his wife, and his own language skills, Max had written Wertheimer's responses in German, Pavlov's responses in Russian, and so on. In order to grade the exam, Dr. Dixon had to enlist assistance from the Foreign Language Department, and with the help of an international student, his graded comments to Max were written in Korean.

There are so many things we will miss about Max. If you were around Max you had to at least smile, if not laugh out loud, doing otherwise was against the rules. He was creative, brilliant, kind, and gentle; a giant of a man, everyone knew when he entered a room. In graduate school and later in other professional arenas, people were in awe of his intellect and wit. Even when surrounded by some of the greatest minds in the world, he was respected for his intelligence. Charlotte Keefe, Professor, Texas Woman's University, was another of Max's dear friends. She described him eloquently, "Max shared his intellect and insights graciously and willingly and usually with uncanny wit. He could 'nail' the essence of a problem with elegance. When I had the privilege of working with him on projects, it was such a pleasure – never a grind because he found humor in even the most trying situations."

Max had so many special gifts. He was a computer whiz, artist, musician, calligrapher, jewelry maker, and photographer. What stood out the most about him was his spirit. His passing is a loss for us all, even those who never knew him. We will never know what additional contributions he could have made or influenced in the field of education, the nation's school children, teachers, professors, and administrators. His statistical expertise was sought after by many, "Hey Max, how do you think I should crunch this data?" "What's the best way to design this study?" He was always willing to take the time to help you figure it out. His own research pertained to a variety of topics including statewide testing and evaluation, hierarchical linear modeling, and complicated cross-cultural issues. In addition to Hispanic educational issues in Texas, Max also considered cross-cultural issues related to Turkey and Korea that undoubtedly touched many lives around the world.

I think we all wonder what people will say about us after we're gone. What will be our legacy? More than anything else, Max was dedicated to his family and to

(*continued on page 6*)

God. His faith was deep and strong. He was happily married to his wife Diane for 25 years – a real accomplishment these days. They were blessed with four children, Jeremy, Max II, Miranda, and Johanna, who are as brilliant, creative, and talented as their dad. Max Martin was my dear friend for almost 20 years. I had the utmost respect and admiration for him both professionally and personally. I am a better person for having known him and I will miss him deeply. Max, from all of us, "Well done, my friend. Well done."

Nancy K. Martin
Associate Dean for Undergraduate Studies
College of Education & Human Development
The University of Texas at San Antonio

Outlier Lies: An Illustrative Example of Identifying Outliers and Applying Robust Models

Karl Ho, University of North Texas
Jimmie R. Naugher, University of North Texas

The presence of outliers can contribute to serious deviance in findings of statistical models. In this study, we illustrate how a minor, typographical error in the data could make a standard OLS model "lie" in the estimates and model fit. We propose robust techniques that are insensitive to extreme, outlying cases and provide better predictions. With implementation examples, we demonstrate how robust technique improves estimations over conventional models based on normality and outlier-free assumptions.

The possibility of outliers is an important consideration when applying regression statistics such as R^2 and the Pearson product moment correlation coefficient (Huber 1981, Hempel *et al* 1986). We provide an example in this article that illustrates how dramatic the influence of only a tiny portion of the data can have on the model estimate and goodness of fit statistics. In the following analysis, we demonstrate that with two outliers included in a data set of 48 observations, only 15% of the variation in the dependent variable is accounted for by the differences on the independent variable ($r = .39$ and $r^2 = .15$, N=48). However, when the two outliers are removed, 48% of the variation is accounted for ($r = .69$ and $r^2 = .48$, N=46).

The data are from a survey of metropolitan colleges and universities conducted by the Office of University Planning at the University of North Texas. The institutions ranged from some with essentially open admissions to those with selective admissions criteria. The independent variable is the *institution's average SAT score for new freshmen* and the dependent variable is the *institution's six-year graduation rate.* As expected, there was a strong linear relationship between the average SAT score for new freshmen and the graduation rates. However, only two outliers can hide this fact in terms of r and r^2 analysis. There are three purposes to this article:

- To illustrate how only two outliers can have a dramatic influence on r and r^2 values.
- To demonstrate that outliers can be identified by visual inspection of the scattergram, provided the difference is extreme enough.
- To point to statistical tools that provide more reliable statistical means to identify outliers than visual inspection alone.

The reported SAT averages ranged from 464 to 1152. The reported graduation rates ranged from 12.0% to 74.4%. The outliers reported the two lowest average SAT scores with relatively high graduation rates, *i.e.*, an SAT of 464 with a graduation rate of 44.1% (near the middle) and an SAT of 598 with a graduation rate of 72.0% (near the top). Institutions were requested to use the total SAT for averages, for which 400 is the lowest possible value. An average SAT of 464 or 598 is not believable. (Probably a clerk recorded either the math SAT or verbal SAT instead of the total SAT. Doubling the two reported SAT values of 464 and 598 yields values that fit well with the graduation rates.)

Figure 1 is based on the 48 cases that include the two outliers. The SAT values and graduation rates are plotted as a graph and the resulting regression line is plotted. Note how the paired values of SAT=464 and graduation rate=44.1 and SAT=598 and graduation rate=72.0 are isolated in the top left corner of the graph. The two points "lie outside" the general pattern formed by the other cases. The R^2 is 0.1523.

Figure 2 is based on 46 cases, with the two outliers excluded. The SAT values and graduation rates as shown in Table 1 are plotted as a graph with the regression line. Note how much better the fit of the regression line with the two outlying cases discarded (R^2 =0.4735).

Identifying and Dealing with Outliers

Apart from visual methods, statistical tools for identifying regression outliers abound. The more commonly known are Mahalanobis distance and Cook's distance. The former measures the distance of a case from the centroid of the remaining cases where centroid is the point created by the means of all variables in a multidimensional space.

$$\text{Mahalonobis distance} = (n - 1)(h_i - 1/n)$$

where n is the number of observations and h_i is the leverage value for ith case derived from the diagonal of the hat matrix $(\mathbf{X'X})^{-1}\mathbf{X'}$.

Cook's distance is another influence measure that reflects the change in the estimates of regression coefficients if the ith case is removed.

$$\text{Cook} = \frac{\text{(h x deleted residual square)}}{\text{(k x residual mean square)}}$$

Figure 3 vividly depicts the outlying observations of the 47th and 48th cases, which Mahalanobis distances are 6.052 and 12.104, respectively, indicating a departure from other cases. Cook's distances for the two cases are 1.039 and 0.664, as compared with the others falling below 0.2.

To circumvent effects of outlying observations, one could remove those cases from the sample, but this sacrifices important information about the outliers.

Table 1. SAT scores and Graduation Rate (GRADRATE)

Case	SAT	GRADRA
1	1152.00	74.40
2	1121.00	69.00
3	1099.00	69.00
4	1069.00	39.00
5	1060.00	68.00
6	1050.00	53.50
7	1044.00	34.00
8	1028.00	41.80
9	1027.00	49.00
10	1026.00	30.00
11	1025.00	47.00
12	1019.00	69.00
13	1009.00	46.00
14	1006.00	50.00
15	1004.00	48.00
16	1000.00	27.00
17	1000.00	45.00
18	998.00	64.00
19	980.00	53.00
20	977.00	34.00
21	968.00	32.00
22	958.00	45.00
23	953.00	46.00
24	927.00	47.00
25	921.00	28.00
26	919.00	44.00
27	918.00	36.00
28	917.00	46.50
29	900.00	50.00
30	892.00	51.00
31	890.00	29.00
32	885.00	25.40
33	876.00	31.00
34	873.00	44.00
35	866.00	41.00
36	857.00	23.00
37	855.00	39.00
38	846.00	37.00
39	831.00	23.00
40	809.00	32.00
41	806.00	12.00
42	799.00	27.00
43	795.00	42.40
44	777.00	41.00
45	760.00	23.00
46	677.00	17.00
47	598.00	72.00
48	464.00	44.10

Deletion of outliers should not be contemplated when the number of cases is substantial. A more positive treatment is to apply Robust Regression techniques that minimize influence of outliers for model estimation.

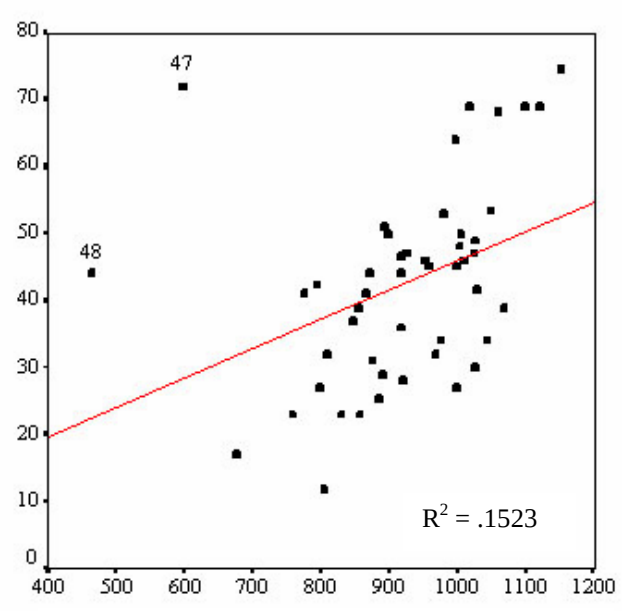


Figure 1. Outliers In: Scattergram of Average SAT and Graduation Rate.

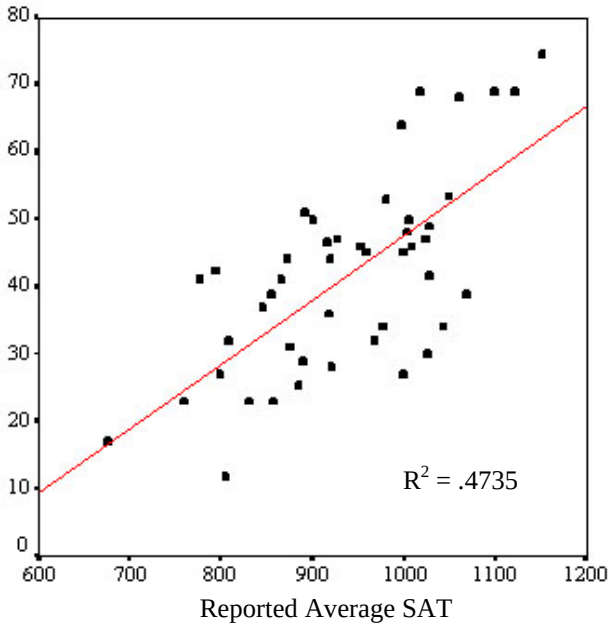


Figure 2. Outliers Out: Scattergram of Average SAT and Graduation Rate.

One of the Robust Regression modeling techniques is based on an MM-estimate computational strategy introduced by Yohai, Stahel and Zamar (1991). The Robust MM Regression method generates highly robust estimates with minimized influence of the outlying cases.

Table 2 lists the model estimates and goodness of fit of the OLS model and Robust MM model using only the SAT score the predict the graduation rate. Notice that the intercept is not statistically significant in the former model. While keeping the two outlying

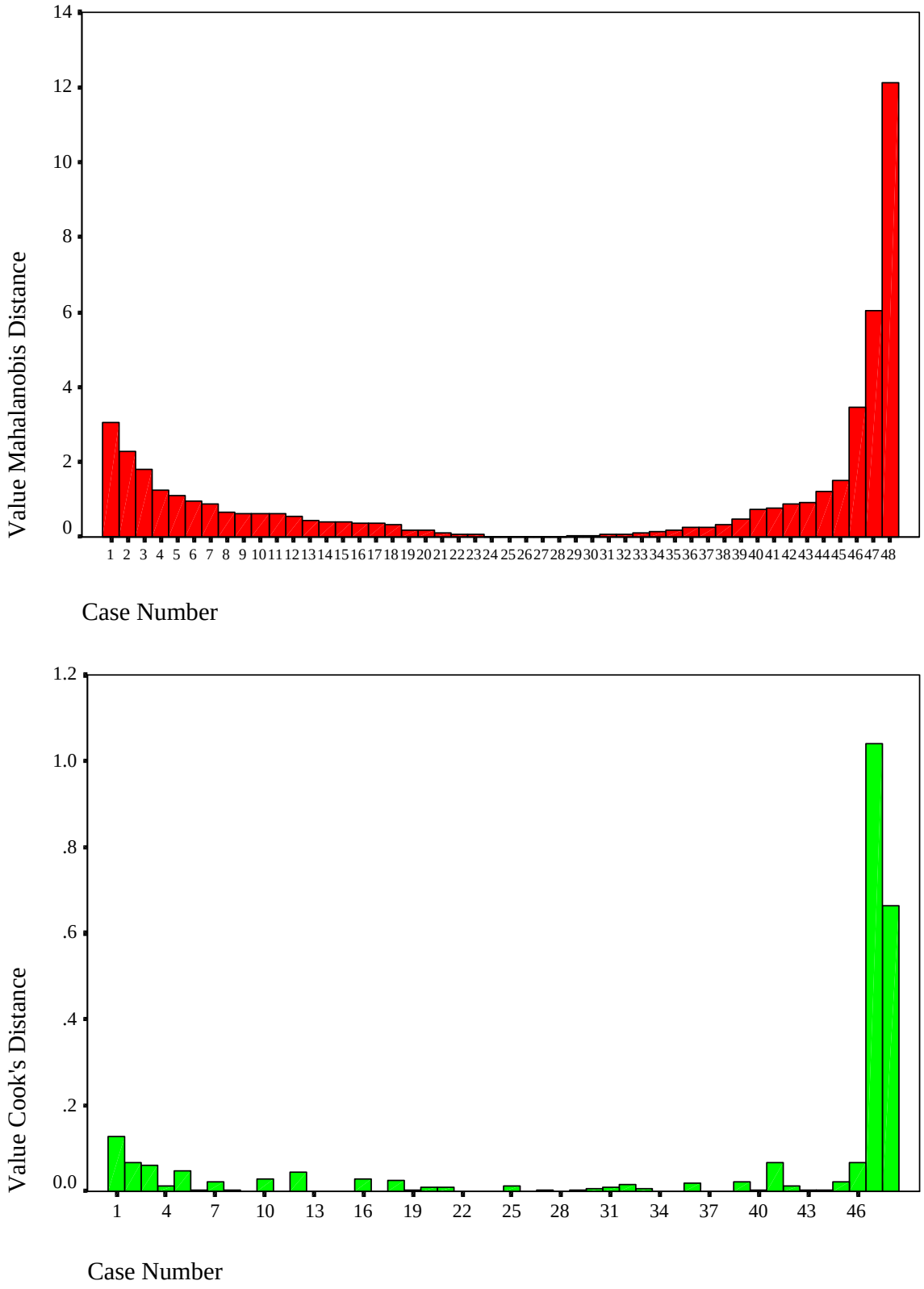


Figure 3. Mahalanobis Distinaces and Cook's Distances.

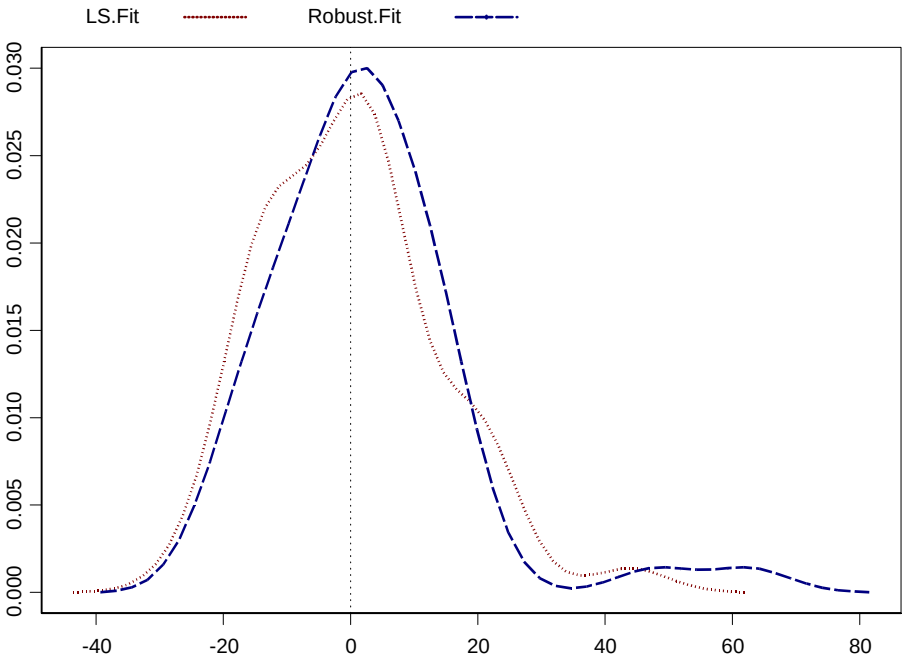


Figure 4. Comparing Densities of Residuals between Robust MM-estimator and Least Square

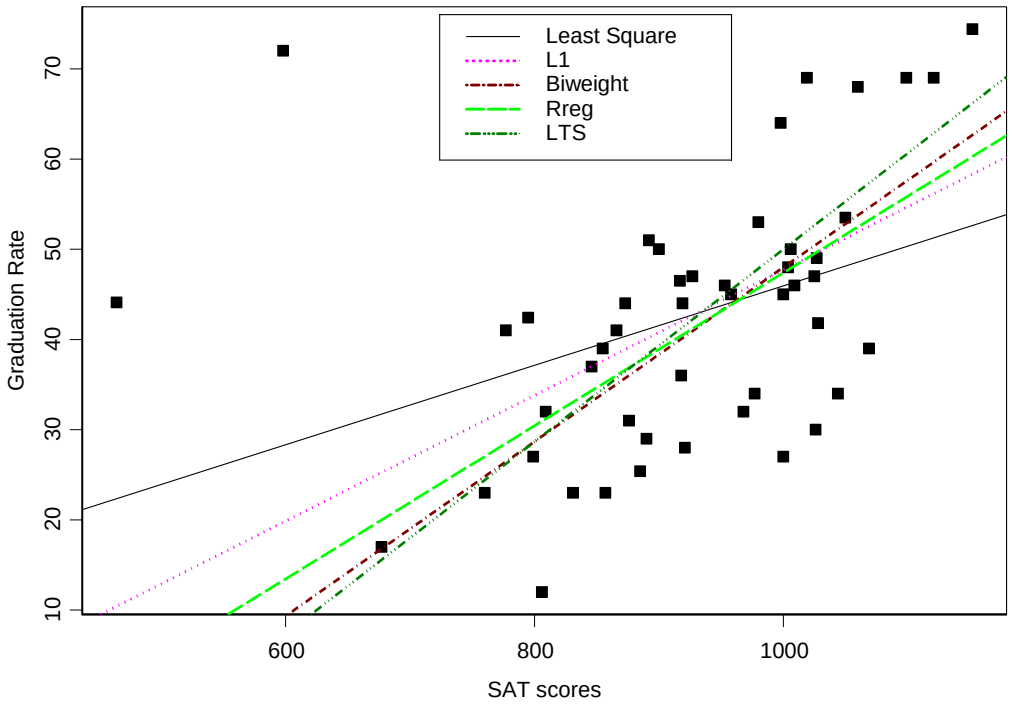


Figure 5. Comparing OLS and Various Robust Estimators

Table 2. Comparison of OLS and Robust Models

OLS	Value	Standard Error	t value	Pr(>|t|)
(Intercept)	1.9170	14.2486	0.135	.893564
SAT	0.0440	0.0153	2.875	.006098
LS Fit :	**0.1523**			
Robust MM				
(Intercept)	-48.2586	16.8244	-2.868	.006209
SAT	0.0960	0.0178	5.382	.000002
LS Fit :	**0.3151**			

cases, namely the 47th and 48th, the Robust MM model does not assume any "manual error" in the data entry but discounts their high influence in modeling the data. The model fit is improved by more than 100 percent.

Figure 4 illustrates how the density of residuals of the robust model is compared to that of the OLS which has bumps on both sides. Comparatively, the robust estimate is well-centered at zero and pushes the outliers farther away to the right.

There are other Robust estimators like Minimum Absolute Residual (L1) Regression, Least Trimmed Squares (LTS), M-estimation(RREG) and Robust Simple Regression by Biweight (Bisquare). Figure 5 demonstrates the relative fit of these robust models compared to OLS. These models can all be implemented using available functions in the S-PLUS 2000 statistical software package (Mathsoft Data Analysis Products Division, 1999).

In conclusion, this article gives a simple illustration of implementing robust models over conventional OLS in the presence of outliers. We demonstrated how outliers can be identified with simple tools and how to deal with data plagued with outlying cases using robust modeling techniques.

References

Hampel, F. Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. 1986. *Robust Statistics: the Approach Based on Influence Functions*. New York: John Wiley & Sons.

Huber, P.J. 1981. *Robust Statistics*. New York: John Wiley & Sons.

Yohai, V., Stahel, W.A. and Zamar, R.H. (1991) A Procedure for Robust Estimation and Inference in Linear Regression, in Stahel, W.A. and Weisberg, S.W., Eds., *Directions in Robust Statistics and Diagnostics, Part II*, Springer-Verlag, New York.

Mathsoft Data Analysis Products Division. 1999. *S-Plus 2000 Guide to Statistics, Volume 1*. Seattle, WA: Mathsoft

(*continued from page 1*)

accomplishment these days. They were blessed with four children, Jeremy, Max II, Miranda, and Johanna, who are as brilliant, creative, and talented as their dad. Max Martin was my dear friend for almost 20 years. I had the utmost respect and admiration for him both professionally and personally. I am a better person for having known him and I will miss him deeply. Max, from all of us, "Well done, my friend. Well done."

Nancy K. Martin
Associate Dean for Undergraduate Studies
College of Education & Human Development
The University of Texas at San Antonio

Problems with Probabilistic Hindsight: A Comparison of Methods for Retrospective Statistical Power Analysis

Jeffrey Kromrey, University of South Florida
Kristine Y. Hogarty, University of South Florida

In contrast to prospective uses of power analysis, retrospective power analysis provides an estimate of the statistical power of a hypothesis test after an investigation has been conducted. The purpose of this research was to empirically investigate the bias and sampling errors of three point estimators of retrospective power and the confidence band coverage of an interval estimate approach. Monte Carlo methods were used to investigate a broad range of research designs and population effect sizes that may be encountered in field research. The results suggest that none of the retrospective power estimation techniques were effective across all of the conditions examined. For point estimates, the "unbiased" and "median unbiased" estimators showed improved performance relative to the plug-in estimator, but these procedures were not completely free from bias except under large sample sizes and large effect sizes (as the statistical power approaches unity). Further the RMSE of these estimates suggests large amounts of sampling error for all three of the point estimators. The interval estimates showed good confidence band coverage under most conditions examined, but the width of the bands suggests that they are relatively uninformative except for large sample and large effect size conditions.

Statistical power analysis is useful from both prospective and retrospective viewpoints. Prospectively, power analysis is used in the planning of inquiry, typically to provide an estimate of the sample size required to obtain a desired level of statistical power under an assumed population effect size, experimental design and nominal alpha level. In contrast, retrospective uses of power analysis involve a consideration of statistical power after inquiry has been completed. This important application of power analysis is somewhat more complicated than the prospective uses.

Two Views on Retrospective Power

Recent literature suggests that retrospective power analysis is conceptualized in two very different forms. Characteristic of one approach, Zumbo and Hubley (1998) and Ottenbacher and Maas (1999) present Bayesian power estimation techniques directed at determining the probability of the null hypothesis being false, given that the null has been rejected, that is *Pr* (H_o=*false*|*rejected* H_o). While this probability is of importance in applied research, it's practical applications appear to be limited because of the unknown proportions of true and false null hypotheses in any field of inquiry (Zumbo & Hubley, 1998). This approach also introduces a different formal definition of "power" than is typically considered in inferential statistics (i.e., power usually represents *Pr* (H_o *will be rejected*|H_o=*false*) which is equal to 1 – β). These two probabilities are often very different. Because this conceptualization of retrospective power is not practical, it will not be further addressed here.

The second approach to retrospective power analysis (Gerard, Smith & Weerakkody, 1998; Steiger & Fouladi, 1997; Brewer & Sindelar, 1987) aims to estimate the statistical power of a hypothesis test after the test has been conducted. That is, information obtained from a particular study may be used to estimate the population effect size, which in turn may be used (in concert with the study's sample size and nominal alpha level) to estimate the power under which the research was conducted. This approach to retrospective power analysis appears to satisfy a practical need in applied research and retains the familiar formal definition of power (i.e., 1 – β). As applied researchers, we have been urged to consider the effect sizes associated with our data (e.g., Kirk, 1996; Harlow, Muliak & Steiger, 1997), in conjunction with the reject/fail-to-reject decisions of our hypothesis tests. The second approach to retrospective power analysis simply extends our use of sample effect sizes to provide estimates of power. However, the estimation of statistical power based on a sample effect size is characterized by considerable controversy.

Estimation Procedures for Retrospective Power

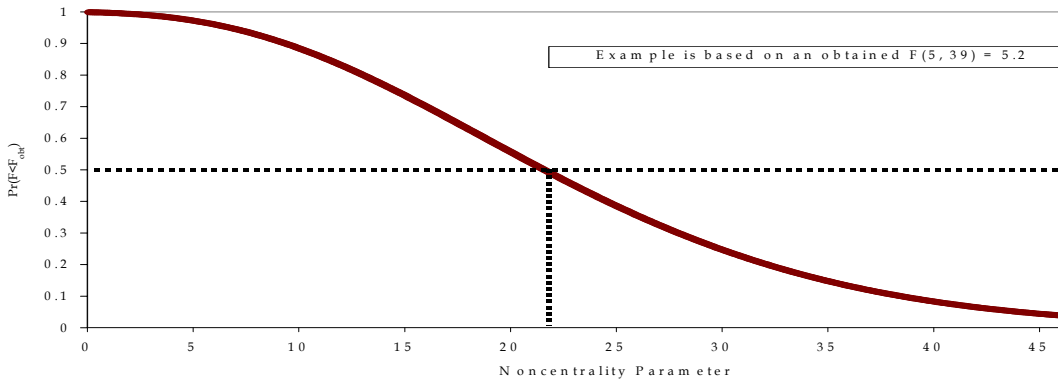
Several techniques for the second approach to retrospective power analysis have been suggested in the literature. Gerard, Smith and Weerakkody (1998) describe three statistics (estimates of noncentrality) that lead to point estimates of retrospective power: a "plug-in estimator" (λ_p), an "unbiased estimator" (λ_{ub}), and a "median unbiased" or "percentile estimator" (λ_{50}) of the noncentrality parameter.

The plug-in estimator simply represents the use of the sample noncentrality parameter (λ_p) as if it were the same as the population parameter. For the F distribution, the sample noncentrality parameter is given by

$$\lambda_p = \nu_1 F$$

where ν_1 = numerator degrees of freedom for the sample *F*, and *F* = obtained sample *F*statistic.

Figure 1
Probability of F < F_{obt} as a Function of Population Noncentrality



The obtained sample noncentrality parameter is then used to estimate the statistical power of the test

$$Power = \Pr(F_{\nu_1, \nu_2, \lambda_p} \geq F_{\nu_1, \nu_2, 1-\alpha})$$

where $F_{\nu_1,\nu_2,\lambda_p}$ = the noncentral F distribution with ν_1 and ν_2 degrees-of-freedom and a noncentrality parameter λ_p,

and $F_{\nu_1,\nu_2,1-\alpha}$ = the $(1-\alpha)$ percentile of central *F*-distribution (i.e., the critical value of *F* with ν_1 and ν_2 degrees of freedom).

The use of λ_p is known to produce biased estimates of power with a distinct positive bias in conditions of low power (Johnson et al., 1995). Johnson et al. suggested an alternative estimator (λ_{ub}) intended to reduced the bias inherent in λ_p. This "unbiased" estimator of noncentrality is given by

$$\lambda_{ub} = \frac{\nu_1(\nu_2 - 2)F}{\nu_2} - \nu_1$$

Although λ_{ub} may provide an unbiased estimate of the population noncentrality, estimates of power derived from unbiased noncentrality estimates are not necessarily unbiased themselves, because power is a nonlinear function of noncentrality (Gerard et al., 1988).

A third point estimate of noncentrality was suggested by Taylor and Muller (1996). This approach (λ_{50}) is reported to underestimate noncentrality 50% of the time and overestimate it 50% of the time (hence, Gerard et al., 1998, refer to the method as "median unbiased"). This method makes use of the cumulative distribution function of *F* and seeks the value of noncentrality for which the obtained value of F in a particular study (i.e., with a given ν_1 and ν_2) is expected 50% of the time (see Figure 1). Because analytical formulae for solving this problem are not available, the value of noncentrality must be obtained by numerical methods (see, for example, Press, Teukolsky, Vetterling & Flannery, 1992).

In contrast to the point estimates suggested by Gerard et al. (1998), Steiger and Fouladi (1997) presented an interval estimation approach based on the earlier work of Hedges and Olkin (1985). This approach provides confidence bands on the noncentrality parameter (noncentrality interval estimates) which subsequently may be used to obtain confidence bands on statistical power. Using logic analogous to that used to obtain the λ_{50} point estimate, the approach involves the inversion of percentiles from noncentral sampling distributions to obtain confidence bands around the noncentrality parameter. That is, instead of seeking the value of noncentrality expected 50% of the time, a 95% confidence band is obtained by seeking the value of noncentrality (λ) for which $\Pr(F_{\nu_1, \nu_2, \lambda} < F_{obt}) = .025$ and the value for which $\Pr(F_{\nu_1, \nu_2, \lambda} < F_{obt} = .975$. This provides a confidence band for noncentrality, the endpoints of which are transformed into the endpoints of a 95% confidence band for statistical power.

Purpose of the Study

Neither the point nor the interval estimation methods for retrospective power analysis have been thoroughly investigated in terms of their operating characteristics. The purpose of this research was to empirically investigate the bias and standard errors of the three point estimators of retrospective power and the confidence band coverage of the noncentrality interval estimate approach. The investigation covered a broad range of research designs and population effect sizes that may be encountered in field research.

Method

A Monte Carlo study was conducted to investigate the bias and standard errors of the three point estimators of retrospective power, and the confidence band coverage of the interval estimation technique. Data were simulated from linear models and sample effect size estimates were used to obtain power estimates. The Monte Carlo study included three factors in the design. These factors were (a) the experimental design simulated, including one factor designs with 2, 4, and 8 levels of the independent variable and three factorial designs (2X2, 2X4 and 3X3), (b) the sample size of the study, with sample sizes ranging from 5 to 100 per cell, including equal and unequal cell sizes, and (c) population effect sizes, with f^2 values (Cohen, 1988) of .01, .02. .15, .35 and .50, as well as a null condition ($f^2 = 0$). The combination of population effect sizes and sample sizes provides conditions with power values ranging from α to nearly 1.00. For each sample generated, the power of the hypothesis test was estimated using the three point estimators and the interval estimate.

The Monte Carlo study was conducted using SAS/IML version 6.12, running on Windows 95 and 98 platforms. The RANNOR random number generator was used to generate normally distributed variables for the observations in each study, and a different seed value for the random number generator was used in each execution of the program. The program code was verified using benchmark datasets.

Fifty thousand replications were conducted for each condition. The use of 50,000 samples provides adequate precision for estimating the relative success of the procedures investigated. For example, the maximum width of a 95% confidence interval around a sample proportion based on 50,000 samples is $\pm$ .0044 (Robey & Barcikowski, 1992).

Results

The results are presented in terms of statistical bias and root mean squared error (RMSE) for the point estimates of power. Statistical bias of the power estimates was estimated as

$$Bias = \frac{\sum_{k=1}^{K}\left(\hat{\theta}_k - \theta\right)}{K}$$

where $\hat{\theta}_k$ = power estimate for the kth sample,
θ = population power, and
K = number of samples simulated.

This statistic represents the difference between the mean sample estimate of power and the true population power for the condition examined.

RMSE of the power estimates was estimated as

$$RMSE = \sqrt{\frac{\sum_{k=1}^{K}\left(\hat{\theta}_k - \theta\right)^2}{K}}$$

This statistic represents the standard deviation of the sample estimates in which deviation is computed from the population parameter rather than from the mean of the sample estimates.

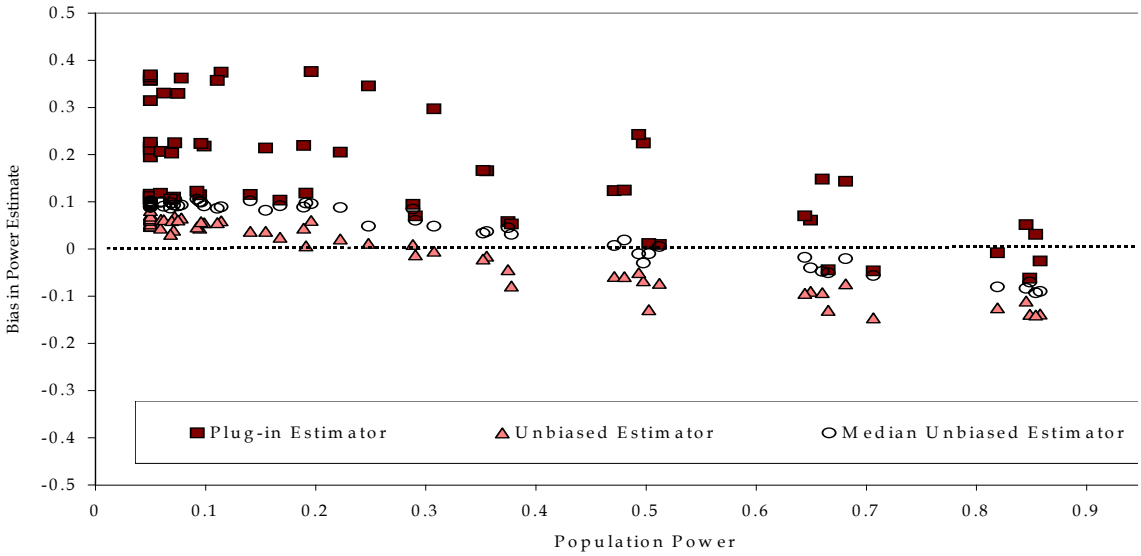
For the interval estimates of power, the proportion of sample confidence bands that contained the parameter were calculated to provide an estimate of the accuracy of the bands. Further, the average width of the confidence bands for each condition was calculated.

To conserve space, results are presented for a subset of the conditions examined (conditions that correspond to Cohen's, 1988, small, medium and large effect sizes in addition to the null condition). Complete results are available from the authors.

Single Factor Designs. Estimates of statistical bias in the point estimates of power for single factor designs are presented in Table 1. Graphs of these bias estimates are provided in Figures 2 and 3. To construct the figures, the population effect size, sample size and number of groups were translated into a population power value which is plotted on the abscissa of each figure. For the null condition ($f^2 = 0$), in balanced designs, all of the estimates evidenced positive bias, with the plug-in estimator presenting the greatest amount of bias (reaching as high as 0.37 for the 8-group design with large samples). Bias evidenced by the plug-in estimator, for a small effect size, was greatest for designs with larger numbers of groups, but the other two estimators did not show such a pattern. The bias in all three of the estimators was reduced as the population effect size increased and many conditions evidenced an underestimate of the power (negative bias). For example, with a medium effect size ($f^2 = .15$), the unbiased estimator evidenced negative bias large as –0.12, with n = 20 in 2-group and 4-group designs. With large samples and a large effect size, all of the estimators converged to the true power (i.e., showing zero bias).

For unbalanced designs, the same pattern was maintained, but the bias estimates were, in general, slightly larger in magnitude. For the null conditions and conditions with a small effect size, a positive bias was evident in most cases, while all of the estimators provided unbiased power estimates for large samples and a large effect size.

Figure 2
Statistical Bias in Point Estimates of Retrospective Power
Balanced Designs



The root mean squared errors (RMSEs) of these point estimates are provided in Table 2. Graphs of these error estimates are provided in Figures 4 and 5. These statistics reflect sampling variability in terms of squared deviations from the population parameter. If a statistic is unbiased, the RMSE is the same as the standard error. Because these statistics reflect sampling error, in many conditions the RMSEs become smaller with larger sample sizes (e.g., for conditions with a large effect size). When estimators are biased, however, the RMSE may not decrease with larger sample sizes. In general the magnitudes of the RMSE associated with these point estimates of retrospective power are quite large for conditions with a small or medium population effect size and small sample size. However, with large samples and large effect sizes, the sampling error is substantially reduced. Further, the magnitude of the RMSE does not appear to be systematically larger with unbalanced designs.

For the interval power estimates, the proportion of confidence bands that contained the true value of power and the confidence interval width are presented in Table 3 and illustrated in Figure 6. For balanced designs, the intervals showed 95% coverage across all non-null conditions, but performance decreased with the unbalanced designs. For the unbalanced designs, confidence band coverage decreased with increasing effect sizes and increasing sample sizes.

As with the RMSE for the point estimates, for both balanced and unbalanced designs, the average width of the confidence bands (Table 3) suggests that the bands are relatively uninformative for small samples and even for large samples if the effect size is small. Only for those conditions with large samples and medium and large effect sizes did the width of the bands become small enough to be considered informative in a practical sense.

Factorial Designs. Estimates of statistical bias in the point estimates of power for factorial designs are presented in Table 4 and illustrated in Figure 7. Consideration of bias for factorial designs must include an examination of row, column and interaction effects. For the null condition ($f^2 = 0$), all of the estimates evidenced positive bias for all three effects, with the plug-in estimator presenting a greater amount of bias for both the column and interaction effects for the 2 X 4 factorial design (approximately .22 across all sample sizes). The greatest amount of statistical bias was seen for the interaction effect for 3 X 3 factorial designs (reaching .26 for all but the smallest sample size). A similar pattern was evidenced for the smallest effect size ($f^2 = .02$) for all but the largest sample sizes. That is, bias in the plug-in estimator, for small effect sizes, was greater for column effects with the 2 X 4 designs and for the interaction effect for both the 2 X 4 and 3 X 3 factorial designs, but the other two estimators did not show such a pattern. Similar to the single factor designs, the bias in all three of the estimators was reduced as the population effect size increased and many conditions evidenced an under-estimate of power (negative bias). For example, with

Table 1. Statistical Bias of Three Point Estimates of Retrospective Power in One Factor Designs.

Balanced Designs

		$f^2 = 0.00$			$f^2 = 0.02$			$f^2 = 0.15$			$f^2 = 0.35$		
Groups	N	λ_p	λ_{ub}	λ_{50}	λ_p	λ_{ub}	λ_{50}	λ_p	λ_{ub}	λ_{50}	λ_p	λ_{ub}	λ_{50}
	5	0.11	0.04	0.09	0.11	0.04	0.09	0.11	0.00	0.09	0.07	-0.07	0.05
	10	0.12	0.05	0.10	0.11	0.04	0.10	0.05	-0.05	0.04	-0.04	-0.14	-0.05
2	20	0.12	0.05	0.10	0.11	0.03	0.09	-0.04	-0.12	-0.04	-0.07	-0.11	-0.07
	50	0.12	0.06	0.10	0.07	-0.02	0.06	-0.06	-0.09	-0.06	0.00	0.00	0.00
	100	0.12	0.06	0.10	0.00	-0.08	-0.01	-0.01	-0.01	-0.01	0.00	0.00	0.00
	5	0.19	0.06	0.09	0.20	0.05	0.10	0.20	0.01	0.08	0.12	-0.07	0.01
	10	0.21	0.06	0.09	0.22	0.06	0.10	0.12	-0.05	0.01	-0.02	-0.13	-0.08
4	20	0.22	0.07	0.10	0.22	0.04	0.09	-0.01	-0.12	-0.08	-0.02	-0.04	-0.03
	50	0.22	0.07	0.10	0.17	-0.01	0.04	-0.01	-0.02	-0.02	0.00	0.00	0.00
	100	0.22	0.07	0.10	0.05	-0.10	-0.05	0.00	0.00	0.00	0.00	0.00	0.00
	5	0.32	0.07	0.09	0.33	0.06	0.09	0.30	0.01	0.06	0.13	-0.10	-0.05
	10	0.35	0.07	0.09	0.36	0.06	0.09	0.15	-0.09	-0.04	-0.01	-0.08	-0.06
8	20	0.36	0.07	0.09	0.37	0.05	0.08	0.00	-0.09	-0.07	0.00	0.00	0.00
	50	0.37	0.07	0.10	0.25	-0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	100	0.37	0.07	0.09	0.06	-0.11	-0.08	0.00	0.00	0.00	0.00	0.00	0.00

Unbalanced Designs

		$f^2 = 0.00$			$f^2 = 0.02$			$f^2 = 0.15$			$f^2 = 0.35$		
Groups	N	λ_p	λ_{ub}	λ_{50}	λ_p	λ_{ub}	λ_{50}	λ_p	λ_{ub}	λ_{50}	λ_p	λ_{ub}	λ_{50}
	5	0.11	0.04	0.09	0.11	0.03	0.09	0.06	-0.03	0.04	-0.02	-0.14	-0.04
	10	0.11	0.05	0.10	0.10	0.03	0.08	-0.03	-0.13	-0.05	-0.18	-0.28	-0.19
2	20	0.12	0.05	0.10	0.08	0.01	0.06	-0.17	-0.26	-0.18	-0.20	-0.27	-0.20
	50	0.12	0.05	0.10	0.00	-0.08	-0.01	-0.18	-0.24	-0.19	-0.03	-0.04	-0.03
	100	0.12	0.06	0.10	-0.11	-0.19	-0.12	-0.04	-0.06	-0.04	0.00	0.00	0.00
	5	0.20	0.06	0.09	0.21	0.06	0.10	0.24	0.05	0.13	0.19	0.01	0.09
	10	0.21	0.06	0.10	0.23	0.07	0.11	0.19	0.03	0.09	0.04	-0.04	0.00
4	20	0.22	0.07	0.10	0.25	0.07	0.12	0.06	-0.03	0.01	0.00	-0.01	-0.01
	50	0.22	0.07	0.10	0.23	0.05	0.10	0.00	-0.01	0.00	0.00	0.00	0.00
	100	0.22	0.07	0.10	0.13	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00
	5	0.32	0.07	0.09	0.35	0.08	0.11	0.38	0.10	0.16	0.21	0.06	0.10
	10	0.35	0.07	0.09	0.40	0.09	0.12	0.23	0.07	0.11	0.02	0.00	0.00
8	20	0.36	0.07	0.09	0.42	0.11	0.15	0.03	0.00	0.01	0.00	0.00	0.00
	50	0.37	0.07	0.09	0.34	0.10	0.14	0.00	0.00	0.00	0.00	0.00	0.00
	100	0.37	0.07	0.10	0.12	0.04	0.05	0.00	0.00	0.00	0.00	0.00	0.00

Note. Estimates are based on 50,000 samples.

Table 2. RMSE of Three Point Estimates of Retrospective Power in One Factor Designs.

Balanced Designs

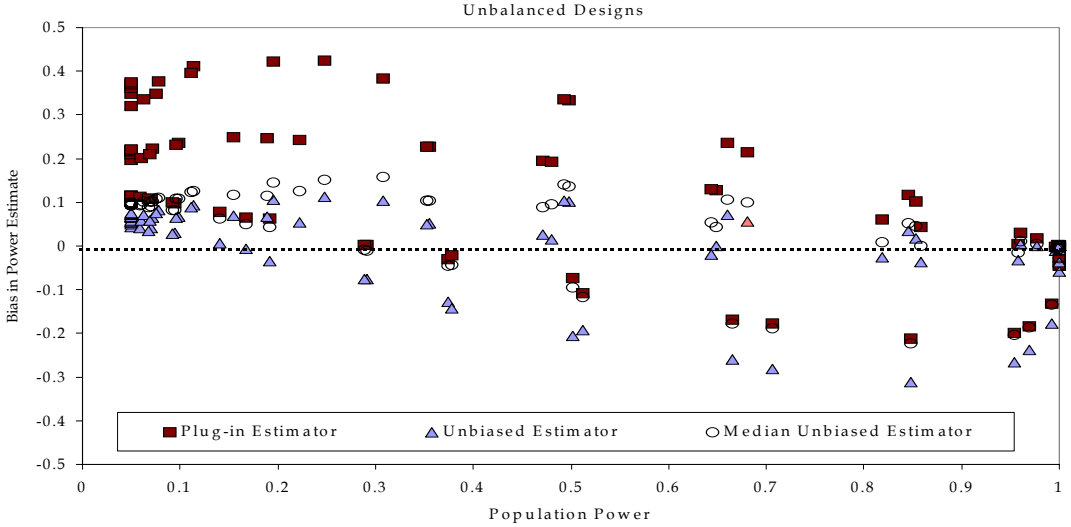
		$f^2 = 0.00$			$f^2 = 0.02$			$f^2 = 0.15$			$f^2 = 0.35$		
Groups	N	λ_p	λ_{ub}	λ_{50}	λ_p	λ_{ub}	λ_{50}	λ_p	λ_{ub}	λ_{50}	λ_p	λ_{ub}	λ_{50}
	5	0.20	0.13	0.19	0.22	0.14	0.20	0.28	0.22	0.27	0.30	0.28	0.29
	10	0.20	0.13	0.19	0.22	0.17	0.22	0.29	0.28	0.29	0.27	0.33	0.28
2	20	0.19	0.14	0.19	0.24	0.20	0.24	0.28	0.33	0.29	0.18	0.23	0.18
	50	0.19	0.14	0.19	0.27	0.26	0.28	0.16	0.20	0.16	0.02	0.02	0.02
	100	0.19	0.14	0.19	0.29	0.31	0.29	0.03	0.04	0.03	0.00	0.00	0.00
	5	0.26	0.14	0.18	0.28	0.16	0.20	0.32	0.23	0.27	0.28	0.30	0.29
	10	0.27	0.15	0.18	0.30	0.18	0.22	0.28	0.30	0.29	0.18	0.29	0.24
4	20	0.28	0.15	0.19	0.32	0.21	0.25	0.19	0.29	0.26	0.05	0.10	0.08
	50	0.28	0.15	0.18	0.30	0.27	0.28	0.04	0.07	0.06	0.00	0.00	0.00
	100	0.28	0.15	0.18	0.24	0.31	0.29	0.00	0.00	0.00	0.00	0.00	0.00
	5	0.38	0.15	0.18	0.39	0.17	0.20	0.38	0.26	0.28	0.22	0.31	0.28
	10	0.40	0.16	0.18	0.42	0.20	0.23	0.23	0.30	0.28	0.06	0.17	0.15
8	20	0.41	0.16	0.18	0.43	0.24	0.26	0.08	0.19	0.17	0.00	0.01	0.01
	50	0.42	0.16	0.18	0.32	0.29	0.29	0.00	0.01	0.01	0.00	0.00	0.00
	100	0.42	0.16	0.18	0.14	0.27	0.25	0.00	0.00	0.00	0.00	0.00	0.00

Unbalanced Designs

		$f^2 = 0.00$			$f^2 = 0.02$			$f^2 = 0.15$			$f^2 = 0.35$		
Groups	N	λ_p	λ_{ub}	λ_{50}	λ_p	λ_{ub}	λ_{50}	λ_p	λ_{ub}	λ_{50}	λ_p	λ_{ub}	λ_{50}
	5	0.20	0.13	0.19	0.21	0.13	0.20	0.24	0.19	0.24	0.28	0.28	0.28
	10	0.19	0.13	0.19	0.21	0.15	0.20	0.27	0.28	0.27	0.34	0.41	0.35
2	20	0.19	0.14	0.19	0.22	0.18	0.22	0.33	0.40	0.34	0.31	0.38	0.32
	50	0.19	0.14	0.19	0.24	0.24	0.25	0.29	0.35	0.29	0.07	0.09	0.07
	100	0.19	0.14	0.19	0.30	0.34	0.31	0.10	0.13	0.10	0.00	0.01	0.00
	5	0.27	0.14	0.18	0.29	0.16	0.21	0.35	0.26	0.30	0.31	0.30	0.30
	10	0.28	0.15	0.19	0.32	0.19	0.23	0.31	0.30	0.30	0.14	0.21	0.18
4	20	0.28	0.15	0.18	0.34	0.24	0.27	0.16	0.23	0.20	0.02	0.05	0.04
	50	0.28	0.15	0.18	0.34	0.29	0.31	0.02	0.03	0.02	0.00	0.00	0.00
	100	0.28	0.15	0.18	0.25	0.28	0.27	0.00	0.00	0.00	0.00	0.00	0.00
	5	0.38	0.15	0.18	0.41	0.19	0.22	0.44	0.30	0.33	0.25	0.25	0.25
	10	0.40	0.16	0.18	0.45	0.22	0.25	0.27	0.26	0.26	0.03	0.06	0.05
8	20	0.41	0.16	0.18	0.48	0.28	0.30	0.04	0.08	0.07	0.00	0.00	0.00
	50	0.42	0.16	0.18	0.38	0.31	0.31	0.00	0.00	0.00	0.00	0.00	0.00
	100	0.42	0.16	0.18	0.14	0.17	0.16	0.00	0.00	0.00	0.00	0.00	0.00

Note. Estimates are based on 50,000 samples.

Figure 3
Statistical Bias in Point Estimates of Retrospective Power
Unbalanced Designs



a medium effect size (f^2 = .15), the unbiased estimator evidenced negative bias of –0.10, for the column, row, and interaction effects with n = 10 for all factorial designs. Once again, with large samples and large effect sizes, all of the estimators converged to the true power (i.e., showing zero bias). For the 2 X 2 factorial designs (in which each effect is tested with a single degree of freedom), trends in bias were similar for all of the power estimates across all effect sizes. However, for the 2 X 4 factorial designs, more striking similarities were witnessed for the column and interaction effects (each tested with three degrees of freedom). While maintaining a similar pattern, in general, the bias estimates were slightly smaller for the row effects than for the column and interaction effects.

The root mean squared errors (RMSEs) of the point estimates are provided in Table 5 and illustrated in Figure 8. An examination of these statistics revealed a considerable amount of error associated with small effect sizes and small samples for all effects examined (i.e. row, column and interaction effects). Substantially less error was evidenced when medium and large effect sizes were paired with larger sample sizes. Additionally, the magnitude of the RMSE did not appear to differ systematically across the row, column, or interaction effects.

For the interval power estimates, the proportion of confidence bands that contained the true value of power are presented in Table 6. For all effects, the intervals showed 95% coverage across all conditions. In general, the average width of confidence bands (Table 6) suggests that these bands are relatively uninformative, that is they provide very little information on true power for small samples and small effect sizes. Only when medium or large effect sizes were paired with large samples sizes, did the width of the bands become small enough to be considered useful.

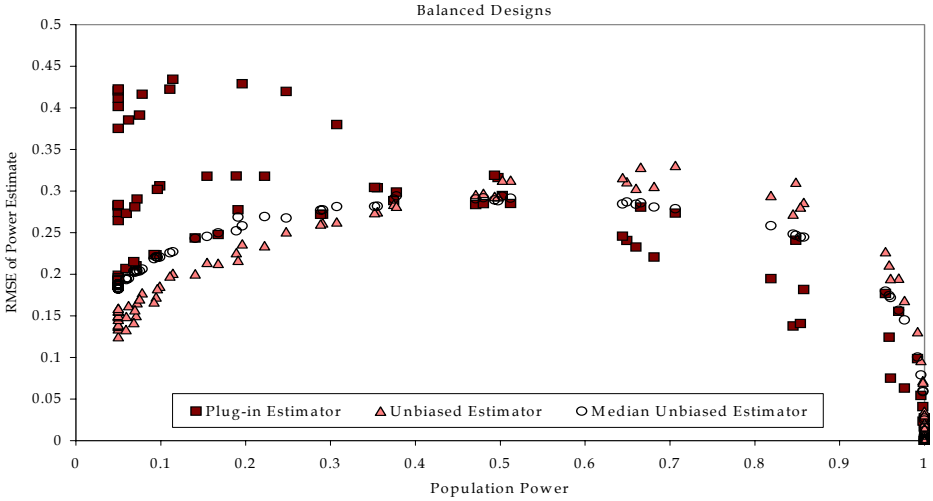
Discussion

The results suggest that none of the retrospective power estimation techniques evaluated were effective across the conditions examined. For point estimates, the "unbiased" and "median unbiased" estimators showed improved performance relative to the plug-in estimator, but these procedures were not completely free from bias except under large sample sizes and large effect sizes (as the statistical power approaches unity). Further, the sampling error in these estimates, reflected in the RMSE, suggests large sampling deviations for all three of the point estimators. These sampling deviations are greatly reduced with large sample estimates of retrospective power.

The confidence band approach suggested by Steiger and Fouladi (1997) provided excellent coverage of the parameter across most of the conditions examined. The coverage problems observed under extreme conditions (i.e., $f^2 = 0$ for both balanced and unbalanced designs, and $f^2 = .35$ with large sample, unbalanced designs) represent research contexts in which power is either zero or very close to one. The calculation of a one-sided confidence interval (e.g., "I am 95% sure that the power is greater than .986") rather than a two-sided band should improve the performance of the confidence bands and may be more useful than a two-sided interval at these extremes.

The coverage results obtained from the confidence band approach suggest that the method appears to be a wise choice (because it is unbiased). However, the width of the resulting confidence bands that provide such excellent coverage were typically so broad that they provided little information about

Figure 4
RMSE of Point Estimates of Retrospective Power
Balanced Designs



the true power of the study. Only with relatively large samples (e.g., n = 100 per cell for one-factor designs) and large effect sizes did the band width become small enough that it appears to be useful for research applications. As with the RMSE associated with the point estimates, the width of these confidence bands reflects the large amount of sampling error that appears to be inherent in retrospective power analysis. For researchers who have the luxury of working with very large samples, these bands appear to be the best approach to power analyses.

Although prospective power analysis is of critical importance in the planning of empirical investigations, retrospective power analysis is important for both the interpretation of research results and the planning of subsequent studies, hence it is a logical extension of the substantive interpretation of sample effect sizes. However, retrospective power analysis has received little attention in the research methods literature. Our results suggest that the currently available methods for retrospective power analysis evidence severe limitations (except for studies with large sample sizes) in terms of statistical bias and large sampling errors. Such results highlight the magnitude of the caveats that should be employed when researchers use retrospective power estimates. Additionally, these results suggest that improved methods of estimation appear to be necessary to supply researchers with an important tool that can be trusted to provide unbiased and precise estimates of retrospective power across conditions typically encountered in applied research.

References

Brewer, J. K. & Singular, P. T. (1987). Adequate sample size: A priori and post hoc considerations. *Journal of Special Education, 21*, 74 - 84.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd Ed.). Hillsdale, NJ: Erlbaum.

Gerard, P. D., Smith, D. R. & Weerakkody, G. (1998). Limits of retrospective power analysis. *Journal of Wildlife Management, 62*, 801 - 807.

Hedges, L. V. & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York: Academic Press.

Johnson, N. L., Kotz, S. & Balakrishnan, N. (1995). *Continuous univariate distributions, Volume 2* (2nd Ed.). New York: Wiley.

Kirk, R. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*, 746-759.

Ottenbacher, K. J. & Maas, F. (1998). How to detect effects: Statistical power and evidence-based practice in occupational therapy research. *American Journal of Occupational Therapy, 53*, 181 - 188.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1992). *Numerical recipes in FORTRAN: The art of scientific computing* (2nd Ed.). New York: Cambridge.

Steiger, J. H. & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. L. Harlow, S. A. Muliak & J. H. Steiger (Eds.), *What if there were no significance tests?* Mahweh, NJ: Erlbaum.

Taylor, D. J. & Muller, K. E. (1996). Bias in linear model power and sample size cacluation due to estimating noncentrality. *Communications in Statistics: Theory and Methods, 25*, 1595-1610.

Zumbo, B. D. & Hubley, A. M. (1998). A note on misconceptions concerning prospective and retrospective power. *The Statistician, 47*, 385 - 388.

Evaluating Univariate, Bivariate, and Multivariate Normality Using Graphical and Statistical Procedures

Tom Burdenski, Texas A & M University

This paper reviews graphical and statistical procedures for evaluating multivariate normality by guiding the reader through univariate and bivariate procedures that are necessary, but insufficient, indications of a multivariate normal distribution. A data set utilizing three dependent variables for two groups provided by George and Mallery (1999) is used to analyze kurtosis and skewness coefficients, Q-Q plots, the Shapiro-Wilk or Kolmogorov-Smirnov statistic, and bivariate scatterplots. A procedure programmed by Thompson (1990) is used to explore multivariate normality by plotting Mahalanobis distances against derived chi-square values in a scatterplot.

Reality is complex. Over time, researchers in the social sciences have become increasingly aware that simple univariate methods comparing an experimental group with a control group on a single dependent variable are inadequate to meet the needs of the complex phenomena that dominate educational and psychological research. In the majority of social science research, two or more dependent variables are necessary, because nearly every effect has multiple causes and nearly every cause has multiple effects. Even when studying a single construct, such as self-concept, it is often helpful to use multiple tools to measure elusive constructs (called "multi-operationalizing").

In a methodological shift that increasingly emphasizes honoring the complexity of reality, Grimm and Yarnold (1995) reported that the use of multivariate statistics in research has accelerated in the last 20 years and that it is difficult to find empirically based research articles that do not employ one or more multivariate analyses. In a comparison of the 1976 and 1992 volumes of the Journal of Consulting and Clinical Psychology (JCCP) Grimm and Yarnold found that the use of multivariate statistics in JCCP increased from 9% to 67% in that 16 year period.

Daniel (1990) noted that multivariate methods usually best honor the reality about which the researcher wishes to generalize. McMillan and Schumacher (1984) compellingly argued against the limitations of viewing the world through an overly-simplified univariate lens:

> Social scientists have realized for many years that human behavior can be understood only be examining many variables at the same time, not by dealing with one variable in one study, another variable in a second study, and so forth. These [univariate] procedures have failed to reflect our current emphasis on the multiplicity of factors in human behavior. In the reality of complex social situations the researcher needs to examine many variables simultaneously. (pp. 269-270)

Thompson (1986, p. 9), stated that the reality about which most researchers strive to generalize is usually one "in which the researcher cares about multiple outcomes, in which most outcomes have multiple causes, and in which most causes have multiple effects." Given this conception of reality, only multivariate methods honor the full constellation of inter-relating variables *simultaneously*.

Experimentwise Error Rates

Whereas "testwise" error rates refer to the probability of making a Type I error for a given hypothesis test, "experimentwise" error rates refer to the probability of having made a Type I error *anywhere* within the study. Inflation of "experimentwise" error rates can be attributed to two factors: (a) the number of dependent variables in the study; and (b) the amount of correlation between the factors--if two factors are perfectly correlated there is no inflation. On the other extreme, very low correlations produce highly inflated "experimentwise" error rates. The Bonferroni inequality can be used to calculate the "experimentwise" error rate when the hypotheses or variables tested using a single sample are perfectly uncorrelated:

$$\alpha_{EW} = 1 - (1 - \alpha_{TW})^K$$

As noted by Thompson (1994):

> ... if three perfectly uncorrelated hypotheses (or dependent variables) are tested using a single sample, each at the α_{TW}=.05 level of statistical significance, the "experiment-wise" Type I error rate will be:
>
> $$\begin{aligned}\alpha_{EW} &= 1 - (1 - \alpha_{TW})^K \\ &= 1 - (1 - .05)^3 \\ &= 1 - (.95)^3 \\ &= 1 - (.95)(.95)(.95) \\ &= 1 - (.9025(.95)) \\ &= 1 - .857375 \\ \alpha_{EW} &= 0.142625\end{aligned}$$

Thus, for a study testing three perfectly uncorrelated dependent variables, each at the α_{TW} = .05 level of statistical significance, the probability is .142625 (or 14.265%) that one or more null hypotheses will be incorrectly rejected within the study. Most unfortunately, knowing this will not inform the researcher as to which one or more of the statistically significant hypotheses is, in fact, a Type I error.

As illustrated by Fish (1988) and Maxwell (1992) using heuristic examples, invoking multiple univariate tests instead of multivariate tests can also lead unwary researchers to fail to identify statistically significant results. The wrong-headed use of the so-called "Bonferroni correction" coupled with use of univariate tests is also inappropriate, because the application (a) severely attenuates power and (b) still does not honor a multivariate reality. Multivariate analyses can detect interaction effects between independent variables that would go undetected if multiple univariate measures were used in place of multivariate measures. Independent variables may have small, but noteworthy effects on multiple dependent variables that add up to an important pattern when examined as a composite, but otherwise appear meaningless in a univariate test (or series of tests) of a single dependent variable.

Assumptions of Multivariate Statistics

Because use of multivariate statistics has become commonplace, it is imperative that researchers understand and honor the central assumptions that guide their use. The first assumption of most multivariate statistics is that the variance/covariance matrices across the *k* groups must be homogeneous (equal); and the second assumption, which is the focus of this paper, is that the interval response variables across the *k* groups must be multivariate normally distributed. The test for homogeneity of variance in multivariate statistics is Box's *M* (Box, 1949; 1954), which is a statistically powerful test of bivariate correlations (unstandardized *r*) that is analogous to the Levene test in univariate analyses. If Box's *M* is favorable, you *do not reject* the homogeneity of variance assumption, which means that you have met the first assumption of multivariate analyses. Box's *M* tests the first assumption, but it is also sensitive to the second assumption of multivariate normality. In other words, if you don't reject the homogeneity of variance assumption, you may have a problem with multivariate normality (see Tabachnick & Fidell, 1983; 1989; 1996 for a detailed elaboration of the homogeneity of covariance assumption).

Univariate Normality

Determining univariate normality is helpful when assessing multivariate normality, because one can do so even with a small sample size ($n < 25$) and because univariate normality is a necessary precondition for multivariate normality (Gnanadesikan, 1977; Johnson & Wichern, 1992). The advantage of proceeding from a univariate to bivariate to multivariate examination of the data is that such a procedure provides useful information on which dependent variables to use before conducting a multivariate analysis. In order to build a foundation for a complete understanding of multivariate normality, a brief review of univariate normality is in order.

Parametric tests require that the sample data be drawn from a population with a known form, most typically the normal distribution, so that at least one population parameter can be estimated from the sample (Munro & Page, 1993). As noted by Bump (1991), the normal curve is determined by a mathematical equation that uses the mean and standard deviation values to determine two additional statistics--skewness and kurtosis. Both statistics are used to assess the normality of a univariate distribution. Skewness refers to the degree of symmetry of the distribution. Kurtosis refers to the shape of the distribution against the normal distribution, by comparing relative height to width. The mean and standard deviation are used to convert the measured scores to z-scores, which are then used to compute the skewness, as explained by Glass and Stanley (1970, p. 91): $K_x = ((\Sigma Z_i^4)/n)$, most researchers and statistical packages, however, apply an additive constant of (-3) so that the skewness will be equal to 0 in a univariate normal distribution."

However, Glass and Stanley (1970) noted that in a univariate distribution, skewness has a very minor effect on alpha or power in ANOVA if the design is balanced (i.e. there are an equal number of observations in each cell) and kurtosis also has a very slight effect on alpha levels and only effects the power of a test when the distribution is platykurtic (flattened as compared to the normal distribution). The severity of the effect of kurtosis on power increases proportionately with the presence of kurtosis in more than one variable.

Graphical and Statistical Tests of Univariate Normality

According to Stevens (1996), one of the most popular graphical methods for testing univariate normality is the normal probability plot or Q-Q Plot (quantile-versus-quantile) in which observations are ordered in increasing degree of magnitude and then plotted against expected normal distribution values. Three additional graphical tests are the box-and-whisker plot, stem-and-leaf plot, and a histogram of the dependent variables. These tests allow a quick and simple means of evaluating the shape of the univariate distribution for each dependent variable. Stevens (1996) recommends that with samples of less than 50 cases, prudent researchers use non-graphical tests such as the chi-square goodness of fit, Kolmogorov-Smirnov test, the Shapiro-Wilk test, *and* an evaluation of the skewness and kurtosis of the distribution to make an evaluation about univariate normality. The Shapiro-Wilk test (Wilk, Shapiro, & Chen, 1968) was developed to detect a wide variety of variations from a normal univariate distribution. The smaller the W value, the greater the departure from normality. As a guideline, Gnandesikan (1977) stated that for $p_{calculated}$ values of 0.1 or higher, normality is a reasonable assumption.

Wilk, Shapiro, and Chen (1968) concluded that for sample sizes under 20, the combination of the skewness and kurtosis coefficients or the Shapiro-Wilk method were most sensitive to detecting extreme non-normality. Stevens (1996) recommended that researchers evaluate unvariate normality by examining the Shapiro-Wilk statistic and examining the kurtosis and skewness coefficients (along with their standard errors) because Shapiro-Wilk has the most power and a review of the skewness and kurtosis can help determine the cause of non-normality whenever it is present. The Shapiro-Wilk test is recommended for samples of less than 25 and the Kolmogorov-Smirnov test is recommended for samples greater than 25. Both the Shapiro-Wilk and the Kolmogorov-Smirnov tests perform an aggregate test of skewness and kurtosis in the univariate case. You do not want to find statistical significance because the null says the distribution is normal and you do not want to reject the assumption of normality.

Bivariate Normality

As noted by Stevens (1996), in addition to establishing univariate normality, two additional characteristics of a normal multivariate distribution are that the linear relationship of any combination of variables is distributed normally, and that all possible subsets of the sets of variables are normally distributed. The relationship between bivariate and multivariate normality is complex. Statistical significance tests like those used in MANOVA require that the distribution of each dependent variable are normally distributed about each of the other dependent variables in any "X_1 and X_2" comparison.

Two distributions that are univariate normal might also be bivariate normal, but just because two distributions are univariate normal does not mean that they will be bivariate normal. In a bivariate comparison, we compare each person's score on two measures, so we are thinking in three dimensions--the X-axis, Y-axis and a third axis to demonstrate frequency of scores. This requirement means that a circular or elliptical pattern will emerge in a scatterplot when examining the correlation of any two dependent variables in a bivariate normal distribution. The narrower the ellipse in the bivariate scatterplot, the greater the correlation between the dependent variables,and subsequently, the greater the likelihood hat the assumption of multivariate normality will hold.

Figure 1 is a graphical representation of a bivariate frequency distribution in two-dimensional form. In this drawing, the viewer is looking down at the distribution from above. The largest concentric circle is the footprint or floor of the bell or mound. The footprint of the bell is not a circle in this example, because the standard deviation for each person on the X-axis is roughly twice as large as the standard deviation on the Y-axis. A series of *contour lines* is used to demonstrate a series of ellipses with varying amounts of distance from the common center, called the *centroid*.

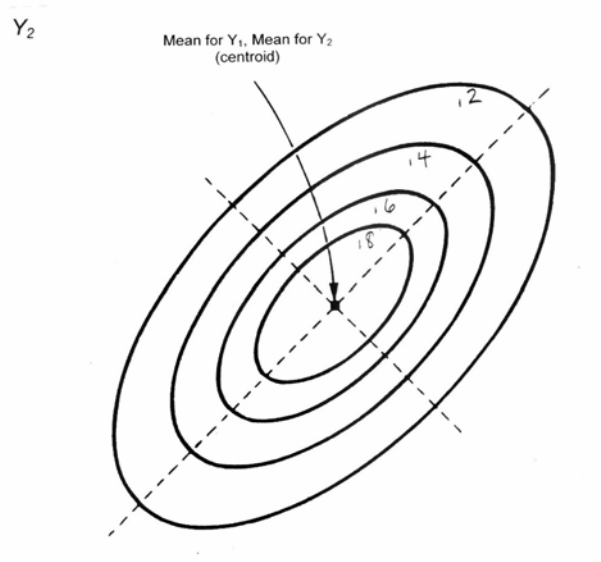


Figure 1. Contour Diagram for a Bivariate Normal Surface

The advantage of drawing the centroid with contour lines is that you can graphically demonstrate the probability that a random bivariate observation (plotted on the X_1X_2 plane) will lie within the elliptical region, which is equivalent to the area under a portion of the normal curve in a univariate distribution of scores (Neter, Kutner, Nachtsheim, & Wasserman, 1996).

Statistical significance testing applies to the bivariate case in terms of the distance from the centroid or Cartesian coordinate for each person on the X and Y axes. The closer the scores aggregate toward the centroid, the greater the chance of being included in the sample because of nearness to the Cartesian coordinate. The first contoured line shows a value of .8 meaning there is an 80% chance of being included in the sample. The last contoured line has a value of .2 meaning that there is only a 20% chance of being included in the sample.

If a group of 400 people is measured in two ways--for example, each person's composite (Verbal + Quantitative) GRE score (X) and self-esteem (Y)--the data can be represented in a bivariate frequency relationship as shown in Figure 2. If we had bivariate normality, the circles would be concentric in a sense. We are comparing two variables, but have three axes. The third axis is height, which graphically shows the frequency of the bivariate scores. In this example, height is a measure of frequency and not a third variable. For each person, there is a pair of scores, a score on X and a score on Y. A bivariate frequency distribution is a picture of the frequency with which different pairs of X and Y scores occur in a group of persons. In Figure 2, a bivariate frequency distribution is displayed for about 400 people on GRE Composite (Verbal + Quantitative) Score (X- axis) and self-esteem (Y-axis). In this example, the highest frequency of scores is a GRE Composite Score of 1000 and a self-

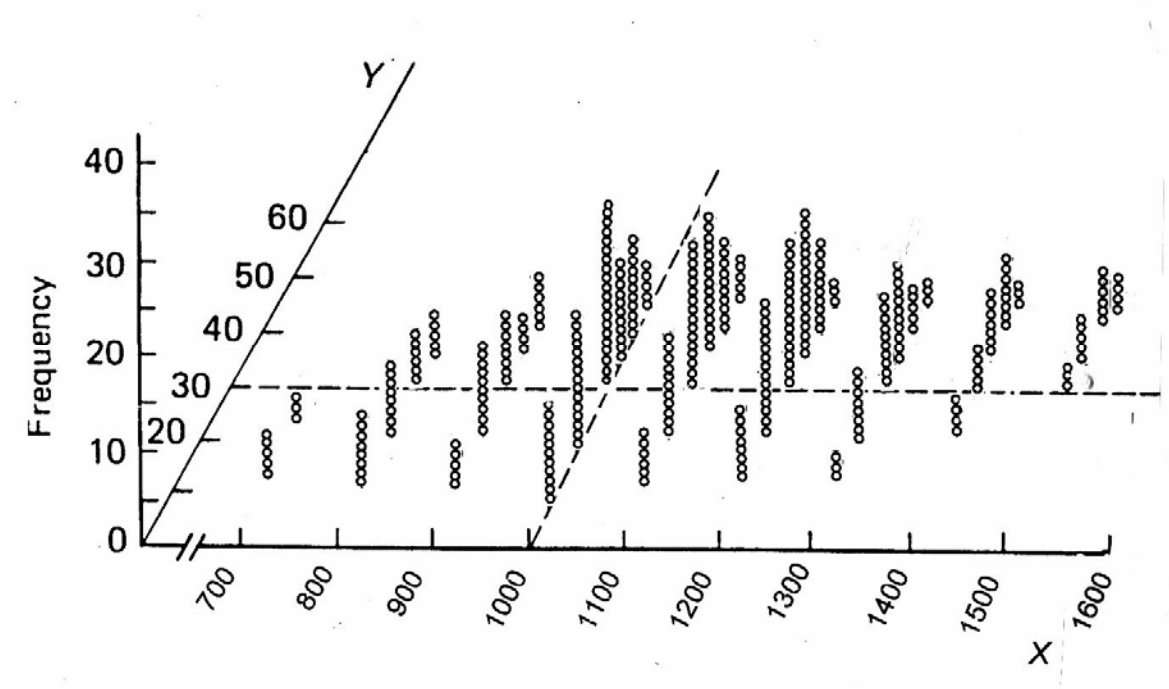


Figure 2. Bivariate Frequency distribution for Persons Measured on Total GRE (X) and Self-Esteem (Y).

esteem score of 30. This point is the Cartesian coordinate for the two sets of scores and also forms the highest point of the distribution of scores. When the height of the line is compared to the vertical scale of frequency, we can determine that approximately 20 persons had a composite GRE score of 1000 and a self-esteem score of 30

A surface or "roof" drawn on the top of a large number of scores in a bivariate frequency distribution takes the shape of a three-dimensional bell or hat as demonstrated in Figure 3. The shape is formed by conceptualizing the one-dimensional bell-shaped normal distribution and stretching it in the X and Y directions and rotating it around its center (i.e. the Cartesian coordinate) in the XY plane. All bivariate normal distributions have the following characteristics:

(a) For each value of X, the distribution of its associated Y value is a normal distribution and vice-versa.

(b) The Y means for each value of X are linear (i.e., they fall on a straight line) and the same is true for the X means for each value of Y.

(c) The scatterplots demonstrate homo-scedasticity--the variance in the Y values is uniform across all values of X and the variance in X values is constant for all values of Y.

If you were to multiply all of the z-scores on the X axis by 2 in Figure 3 and place those scores on the Y axis, the base of the three-dimensional bell will be an ellipse instead of a circle because the Y scores will be twice as spread out as the X scores. However, a non-circular base can still be normal because a multiplicative constant of two will not change the skewness, kurtosis, or mean of zero.

The shape of the mound or hat is determined by the amount of correlation between the two variables. If both dependent variables are expressed in standard deviation units, the more correlated the variables, the narrower the mound or hat because correlation causes the probability to concentrate along a line (see Figure 4; r = .8). In the extreme case that dependent variable X_1 is completely correlated with dependent variable X_2, all points would be exactly on the regression equation, the standard deviation for X_1 and X_2 would be equal to zero and the "contour" would all be straight lines with no areas.

Furthermore, if the distribution is bivariate normal, any plane perpendicular to the X_1X_2 plane will cut the surface into a normal curve and a plane parallel to the X_1X_2 plane will cut in an ellipse. The bivariate normal distribution has the property that the regression of X_1 on X_2 is linear. Therefore, if we have a bivariate normal distribution, we know that if we trace the means of X_2 for each X_1, the result will be a straight line. It does not necessarily follow, however, that if the regression is linear, the joint distribution is bivariate normal.

Multivariate Normality

For a data set of two or more dependent variables, all of the variables must be univariate normal and all possible pairs of the variables must also be normal as

necessary but insufficient conditions for multivariate normality. The mathematical model that serves as the basis for MANOVA and other multivariate techniques is based on the multivariate normal distribution. This means that both the sampling distributions of the means of dependent variables in each cell are normally distributed as are the linear combinations of dependent variables. The central limit theorem states that for large samples, the sampling distribution of means in the univariate case will approach normality. Mardia (1971) demonstrated that MANOVA is robust to modest violation of normality if the violation is caused by skewness rather than outliers.

In some instances, researchers can examine multivariate outliers by simply examining z-scores and looking for extreme scores on each dependent variable. However, this technique does not identify a set of scores for a person that are slightly deviant on several variables. Fortunately, a statistic called Mahalanobis distance (D^2) can be used to detect scores that deviate from the mean (above or below) for a group of dependent variables as a set. Detecting multivariate outliers from a set of dependent variables is a much subtler process than detecting univariate or bivariate outliers.

The Mahalanobis distance is the distance of a case from the centroid where the centroid is the point defined by the means of all the variables taken as a whole. The Mahalanobis distance demonstrates how far an individual case is from the centroid of all the cases for the predictor variables. When the distance is great, the observation is an outlier. According to Krzanowski (1988) and Stevens (1996), the Mahalabonis distance is accepted by researchers as the measure of distance between two multivariate populations and it is independent of sample size. The Mahalanobis distance can be written in terms of the covariance matrix **S** as:

$$D_i^2 = (X_i - x)' \mathbf{S}^{-1} (X_i - x),$$

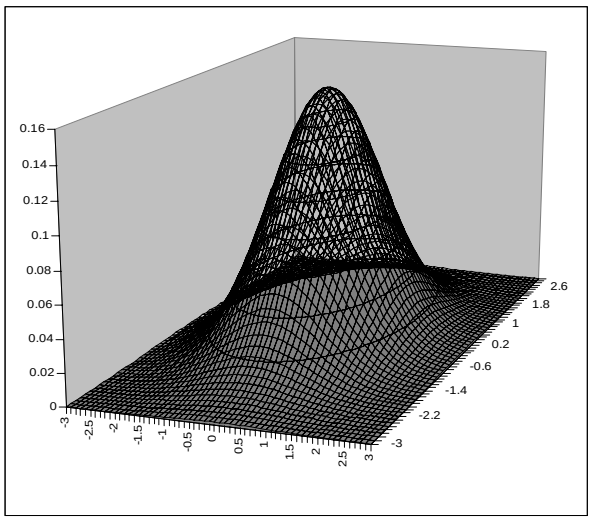
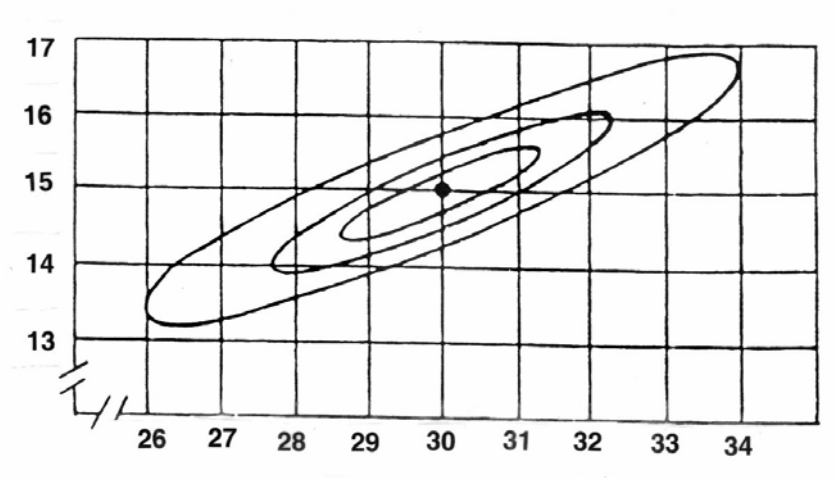


Figure 3. Bivariate Normal Distribution

where D_i^2 is the Mahalanobis distance for a given individual, **S** is covariance matrix with variances on the diagonal and covariances off the diagonal. The rank for S is the number of rows and columns for the covariance matrix, which is 3 x 3, if there are three dependent variables.

The assumption of MANOVA, for example, is that in each group, multivariate normality holds regarding the dependent variables, so if there are a total of 105 cases (as in the heuristic example below) with 64 cases in the female group and 41 cases in the male, both have to have multivariate normality. In group 1, there are three interval variables and the rank of the correlation matrix is 3 x 3. X_i is the composite of three scores of a given individual with a rank of 3 x 1. Person #1 has three scores with one column. The matrix of means also has a rank of 3 X 1 (three means with one column) which yields a product of 3 X 1 and is not conformable to 3 x 3. The transpose (') notation means you flip the



Mean for $X_1 = 30$, $SD_1 = 2$. Mean for $X_2 = 15$, $SD_2 = 1$. $r = 0.8$

Figure 4. Elliptical Bivariate Normal Distribution for 2 Variables with Dissimilar Standard Deviations and Means

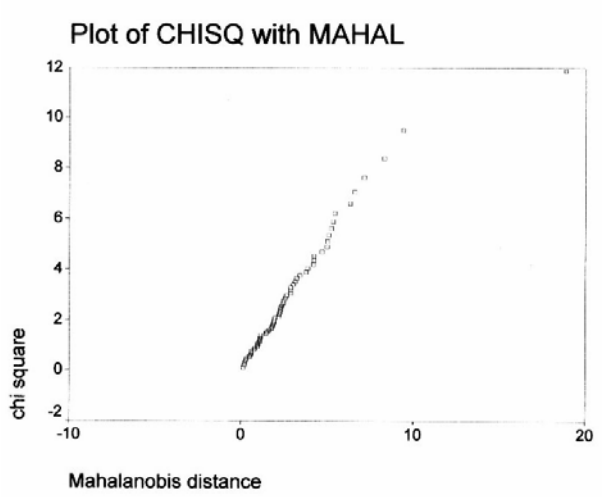


Figure 5. Scatterplot of Chi-Square with Mahalanobis Distance for 64 females without transforming or deleting scores.

3 x 1 and it becomes 1 x 3. The right most part of the matrix is also a 3 x 1 but it does not have a transpose symbol, so it is not flipped on its side.

From the formula, the Mahalanobis distance is descriptive of how far each case's set of scores is from the group means adjusting for correlation of the variables (in the example, a measure of the distance of the each person from the group means adjusted by how correlated the three variables are). In Figure 5, the smallest Mahalanobis distance is for participant #32 because each of the three scores (3.0, 6.1, and 9.8, respectively) is closest to the mean for each variable (2.89; 6.23; and 10.3, respectively).

Having correlated dependent variables is commonplace in social science research. The correlation of dependent variables must be taken into account when calculating the Mahalanobis distance because deviations from the means of two highly correlated dependent variables are partially redundant whereas the deviations from the mean for two highly uncorrelated dependent variables are not redundant. More concretely, say in a set of three dependent variables all with a standard deviation of 5, that the mean of X_1 is 10, the mean of X_2 is 11 and the mean of X_3 is 2, X_1 is highly correlated with X_2 but X_1 is highly uncorrelated with X_3 and X_2 is highly uncorrelated with X_3. If person #1 has a score one standard deviation above the mean on X_1 (X_1=15) and X_2 (X_2=16) and scores at the mean of X_3 (X_3=7), that person will have a smaller D^2 than person #2 who scores at the mean on X_1 (X_1=7) and one standard deviation above the mean on X_2 (X_2=16) and X_3 (X_3=7). The D^2 for person #1 includes redundant distance from the means because the scores on X_1 of 15 and X_2 of 16 are very similar. In a sense, X_1 and X_2 are measuring the same thing, so the deviation from the means is due in part to similarity in the variables. Person #1 will have a lower D^2 because the deviation from the means is redundant whereas the D^2 for person #2 will be much greater because the Mahalanobis distance is not due to distance from similar means of the variables but rather to substantial distance from *dissimilar* means (X_1=10; X_2=16; X_3=7).

There are two evaluations to be done when examining the Mahalanobis distance by chi-square scatterplot--the first is whether or not the points form a straight line or not. If the points on the scatterplot form a straight line, you have multivariate normality. The second consideration is whether or not there are anomalous persons with scores on the scatterplot that are a noteworthy distance from the centroids. You can have a perfectly straight line and still have outliers in the data set, but it is rare to have a person whose scores are outlying on all of the dependent variables in a data set. Before eliminating outliers, a prudent researcher will examine whether or not the extreme score on the multivariate scatterplot is due to an anomalous score on one dependent variable by examining each univariate distribution before eliminating the person from the data set. If only one score is anomalous, it is more prudent to transform the score on that variable rather than eliminate valuable information from the analysis, or to eliminate that variable from the data set.

Evaluating Univariate Normality: A Heuristic Example

To make the discussion about testing bivariate and multivariate normality more concrete, a data set developed by George and Mallery (1999) will be analyzed using SPSS version 8.0 to test the distribution of scores for 64 female and 41 male students taught by the same professor in three sections of a course. The three dependent variables in this analysis are each student's GPA previous to taking the course (PREVGPA), final exam grade (FINAL) and total points for the course (TOTAL). In such a data set, it might be interesting to examine the differences between males and females (an independent variable with two levels) on all three dependent variables--previous GPA, final exam grade, and total points in the course. The SPSS syntax for the female group (n = 64) appears in Appendix A and the syntax for the male group (n = 41) appears in Appendix B. For the sake of brevity and clarity, univariate normality will be assumed and only the bivariate and multivariate output from the female group will be analyzed in detail in this paper.

As noted by Marascuilo and Levin (1983), multivariate normality is a requirement for utilizing the statistical inference procedure that is the basis of all "OVA" designs. The test for univariate normality for the grades data for the female group was done by using the MULTINOR program developed by Thompson (1990) on SPSS 8.0 (Appendix A). The MULTINOR program generates graphical and non-graphical analyses of the distribution of each dependent variable separately.

Bivariate Normality

If the three dependent variables displayed univariate normality (bearing in mind that univariate normality is a necessary, but insufficient foundation for multivariate normality), the next step would be to examine the bivariate correlations between each of the dependent variables. You can attain univariate normality, but fail to demonstrate bivariate normality, which examines each pair of variables--PREVGPA with FINAL, PREVGPA with TOTAL and FINAL with TOTAL. This was done in this example by using the MULTINOR program (Appendix B) by requesting scatterplots and noting elliptical patterns for the three possible combinations of variables. In Figure 6, the scatterplot for each possible pair reveals a clear ellipitical pattern between FINAL and TOTAL, but the scores in the scatterplots for PREVGPA with FINAL and PREVGPA with TOTAL are widely scattered and are thus *not* bivariate normal. When the pattern of the scores in a bivariate plot are less clear, researchers can examine the percentage of scores that converge around the centroid (e.g., 80%, 60%, 40%, 20%, 10%) as a guide to deciding whether or not an elliptical pattern is displayed.

At this stage of the analysis, a prudent researcher might stop and consider replacing PREVGPA with another dependent variable or go back and transform the scores in each of the univariate distributions to make them more normal. As noted earlier, Tabachnick and Fidell (1996) recommended that researchers start by taking the square root of the scores, but the scores can also be squared, or the natural log or log-ten (LG10) can be used:

> ...transformations may improve the analysis, and may have the further advantage of reducing the impact of outliers. Our recommendation, then, is to consider transformation of variables in all situations unless there is some reason not to. If you decide to transform, it is important to check that the variable is normally or near-normally distributed after transformation. Often you need to try one transformation and then another until you find the transformation that produces the skewness and kurtosis values nearest zero, the prettiest picture, and/or the fewest outliers. (p. 82)

After transforming the univariate distributions, the bivariate distributions could be examined again to determine if the three pairs of variables have become bivariate normal due to the univariate transformation of scores. For this set of scores, four data transformations were conducted: (a) square root of scores (Figure 7), (b) squared scores (Figure 8) (c) natural log (Figure 9), and (d) log 10 (Figure 10). In none of these transformations did the bivariate relationships between PREVGPA and TOTAL or PREVGPA and FINAL become bivariate normal. Because PREVGPA appeared to be the problematic DV, a decision was made to create a new DV that was comprised of the sum of the quiz grades in the course. This new DV was named QUIZTOT and a new evaluation of univariate, bivariate, and multivariate normality was conducted as before. The syntax commands for the new variable are shown in Appendix D. Figure 11 shows that the variable QUIZTOT has a bivariate normal relationship with both FINAL and TOTAL and is a big improvement over the variable PREVGPA.

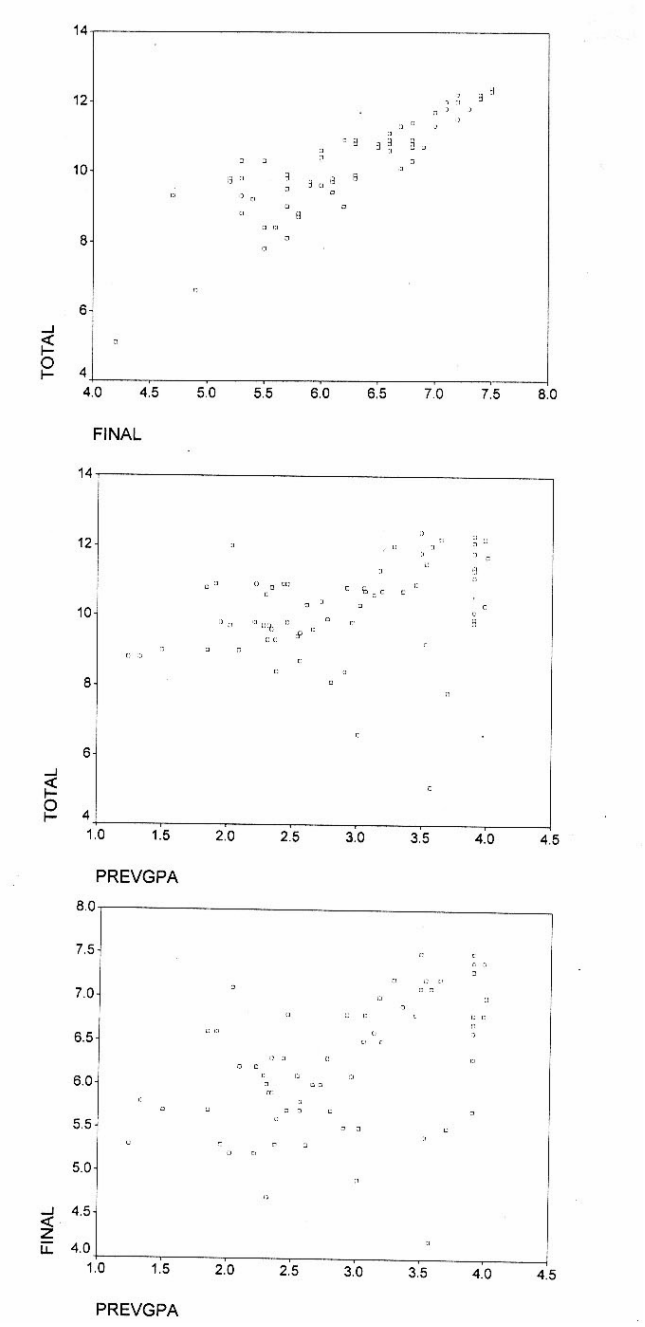


Figure 6. Bivariate Scatterplots of PREVGPA, TOTAL, and FINAL.

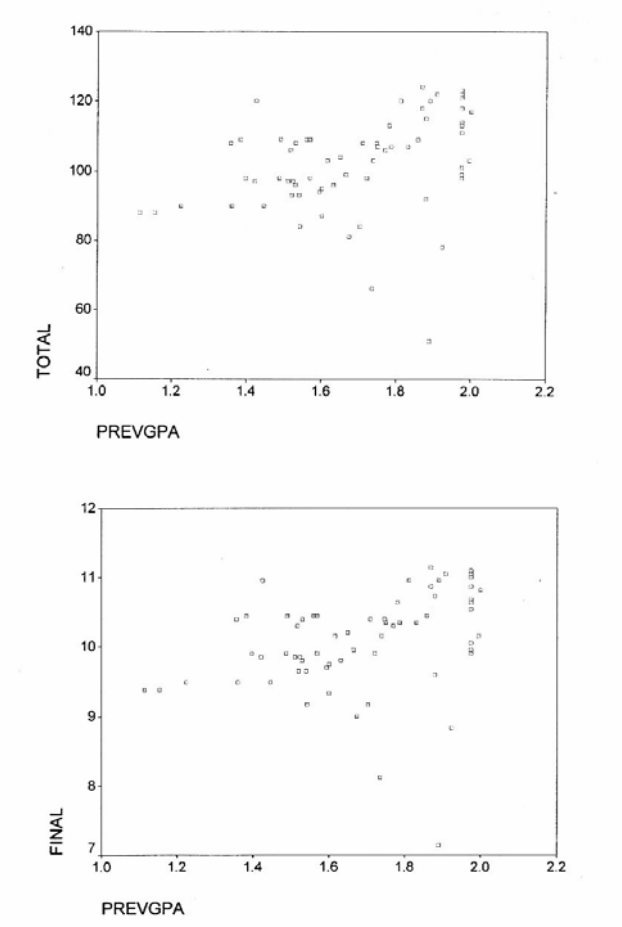


Figure 7. Bivariate Scatterplots of PREVGPA with TOTAL and FINAL (square-root transformation).

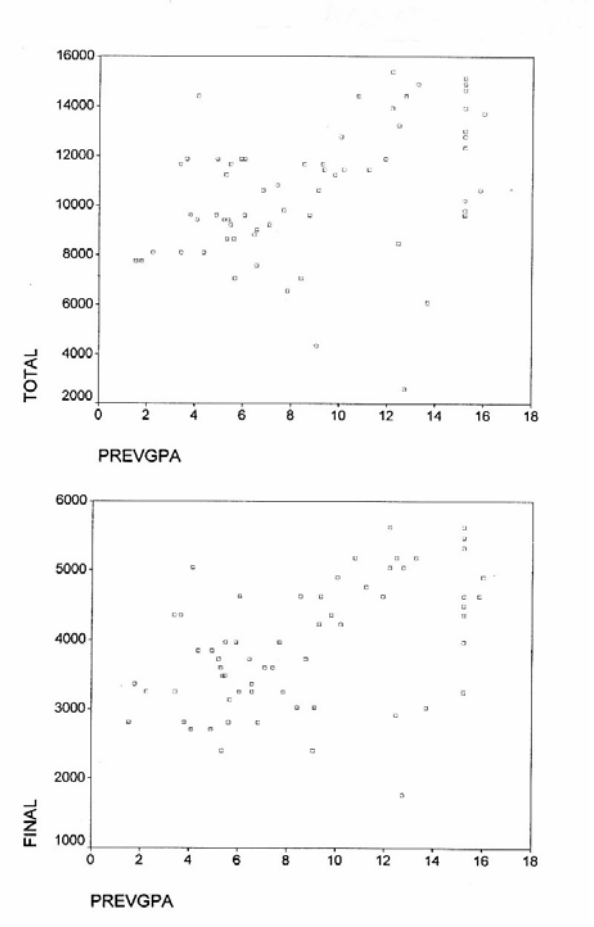


Figure 8. Bivariate Scatterplots of PREVGPA with TOTAL and FINAL (squared-score transformation).

Multivariate Normality

Assuming that both univariate and bivariate normality are attained after transforming the univariate scores or replacing a dependent variable (as done in this example), the third level of assessment is to examine the Mahalanobis distance by chi-square scatterplot to assess multivariate normality. As noted earlier, the Mahalabonis distance is accepted by researchers as the measure of distance between two multivariate populations and it is independent of sample size (Krzanowski, 1988; Stevens, 1996). If we examine the scatterplot of Mahalanobis distance (D^2) values with chi-squares (Thompson, 1990) for this data set in Figure 12 we can see that we have a fairly straight line, which suggests multivariate normality. The second issue is the presence of outliers. This scatterplot has one extreme score in the upper right hand corner that is well off the line with an approximate D^2 score of 62 and a chi-square score of about 12. If we look at the listing of Mahalanobis distances which are ranked from lowest to highest in Figure 16, we can determine that the outlier is case #61 and the D^2 is more than four times larger than the next largest D^2 (case #36). Because case #36 in turn is twice as large as the next largest D^2 (case #45), both case #61 and #36 can be considered outliers. Again, assuming univariate and bivariate normality has been demonstrated, because we have multivariate normality except for two outliers, we can remove or transform the outliers and then look at the univariate and bivariate relationships again because removal of the extreme scores will change the means for both variable X and variable Y, which means that the Mahalanobis distance for each variable will change. If after examining the raw data for case #36 and #42, we discover that they both had very high quiz scores (QUIZTOT) and very low scores on the FINAL, we might call these two students and ask why they did so poorly on the final exam. If we learn that they both had the flu the day of the exam, but took the exam anyway, we might delete their scores from the data set because their illness likely produced "fluky" or abnormal scores (i.e. high quiz scores and low final exam scores). Figure 13 shows the Mahalanobis distance and chi-square values for this data set after the outliers are re-

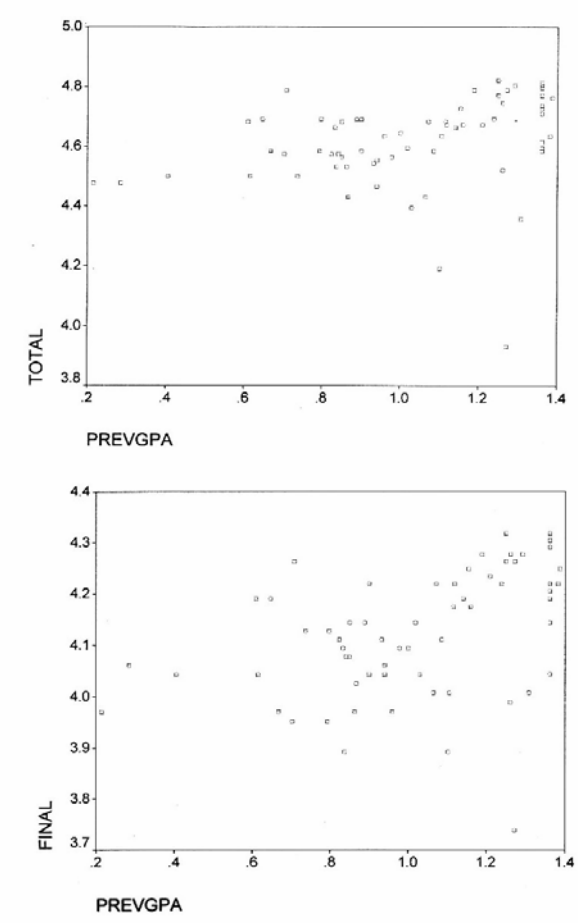


Figure 9. Bivariate Scatterplots of PREVGPA with TOTAL and FINAL (natural log transformation).

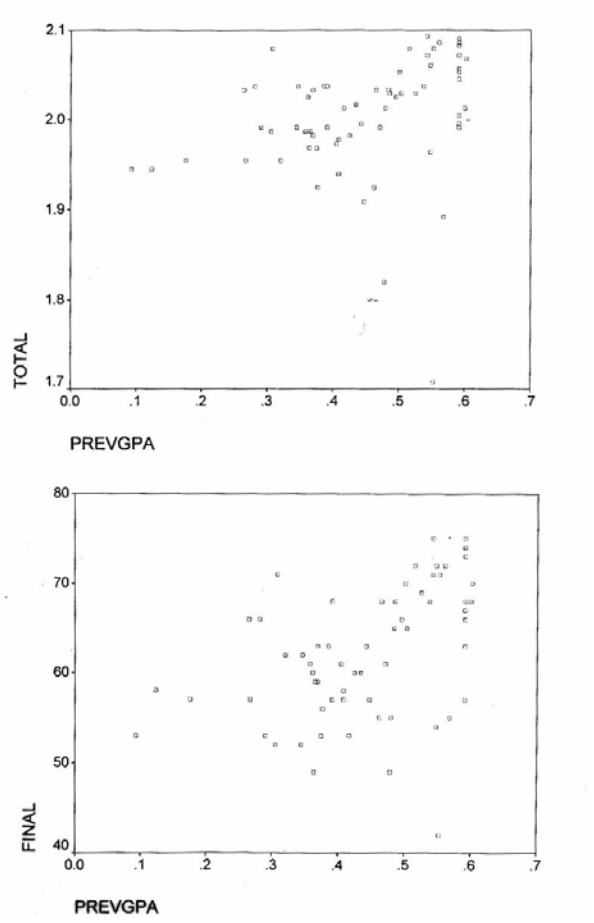


Figure 10. Bivariate Scatterplots of PREVGPA with TOTAL and FINAL (log-10 transformation).

moved. Note that while the line appears to become less straight, in actuality the scale for the Mahalanobis distance is being reduced from 70 units to 12 units, thus showing more precisely the linear relationships between the two variables.

An alternative to the stair-step approach of examining the univariate, bivariate, and multivariate normality of the proposed dependent variables in sequence for the multivariate analysis is to plot the Mahalanobis distance against the chi-square values straight away--if you get a straight line, you can stop there because multivariate normality subsumes univariate and bivariate normality. However, plotting Mahalanobis distance against chi-square is only useful with samples greater than 25. If you fail to obtain a straight line, you can remove scores when you can justify doing so, or transform an individual's scores or a set of scores.

References

Box, G.E.P. (1949). A general distribution theory for a class of likelihood criteria, *Biometrika, 36,* 317-346.

Box, G.E.P. (1954). Some theorems on quadratic forms applied in analysis of variance problems: II. Effect of inequality of variance and of correlation between errors in the two-way classification. *Annals of Mathematical Statistics, 25*, 484-498.

Bump, W. (1991, January). *The normal curve takes many forms: A review of skewness and kurtosis*. Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio. (ERIC Document Reproduction Service No. ED 342 790)

Daniel, L.G. (1990, January). *Use of structure coefficients in multivariate educational research: A heuristic example.* Annual Meeting of the Southwest Educational Research Association, Austin, TX. (ERIC Document Reproduction Service No. ED 315 451)

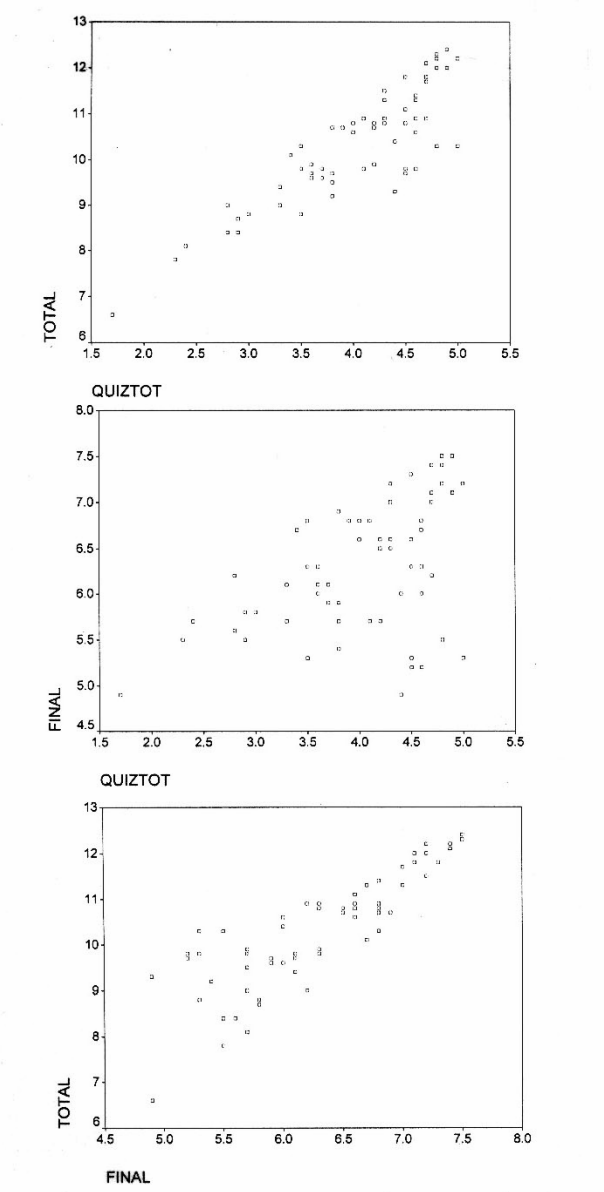


Figure 11. Bivariate Scatterplots of QUIZTOT with TOTAL and FINAL.

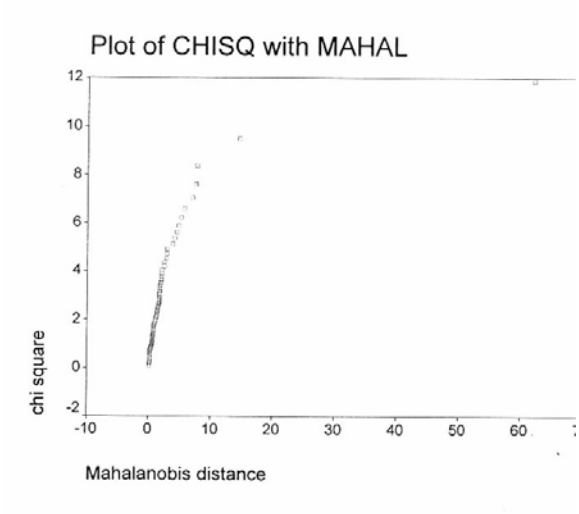


Figure 12. Scatterplot of Chi-Square with Mahalanobis Distance for 64 females after replacing PREVGPA with QUIZTOT.

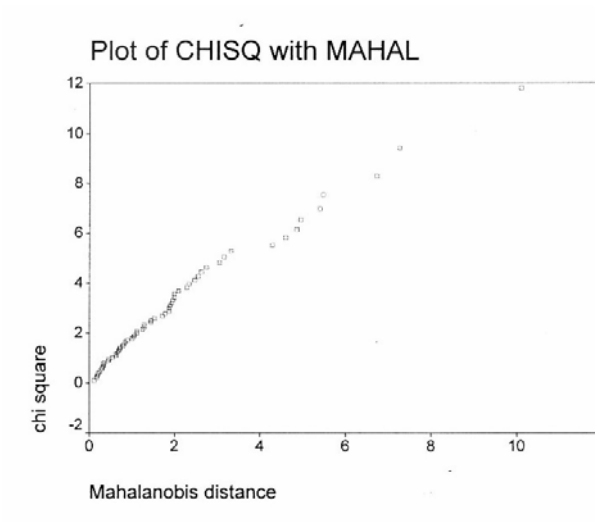


Figure 13. Scatterplot of Chi-Square with Mahalanobis Distance for 64 females after replacing PREVGPA with QUIZTOT and deleting two outliers.

Fish, L. (1988). Why multivariate methods are usually vital. *Measurement in Evaluation and Counseling and Development*, *21*, 130-137.

George, D., & Mallery, P. (1999). *SPSS for WINDOWS step by step*. Boston: Allyn & Bacon.

Glass, G.V., & Stanley, J.C. (1970). *Statistical methods for education and psychology*. Englewood Cliffs, NJ: Prentice-Hall.

Gnandesikan, R. (1977). *Methods for statistical analysis of multivariate observations*. New York: Wiley.

Grimm, L.G., & Yarnold, P.R.. (1995). *Reading and understanding multivariate statistics.* Washington, DC: American Psychological Association.

Johnson, N., & Wichern, D. (1988). *Applied multivariate statistical analysis* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Krzanowski, W.J. (1995). *Recent advances in descriptive multivariate analysis.* Oxford: Clarendon

Marascuilo, L.A., & Levin, J.R. (1983). *Multivariate statistics in the social sciences: A researcher's guide*. Monterey, CA: Brooks/Cole.

Mardia, K.V. (1971). The effect of non-normality on some multivariate tests and robustness to non-normality in the linear model. *Biometrika*, *58*, 105-121.

Maxwell, S. (1992). Recent developments in MANOVA applications. In B. Thompson (ed.), *Advances in social science methodology* (Vol . 2, pp. 137-168). Greenwich, CT: JAI Press.

McMillan, J.H., & Schumacher, S. (1984). *Research in education: A conceptual approach.* Boston: Little, Brown.

Munro, B., & Page, E. (1993). *Statistical methods for health care research* (2nd ed.). Philadelphia: J. B. Lippincott.

Neter, J., Kunter, M., Nachtsheim, C., & Wasserman, W. (1996*). Applied linear statistical models* (4th ed.). Chicago: Irwin.

Stevens, R. (1991). *Applied multivariate statistics for the social sciences* (2rd ed.). Mahwah, NJ: Erlbaum.

Stevens, R. (1996). *Applied multivariate statistics for the social sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Tabachnick, B.G., & Fidell, L.S. (1983). *Using multivariate statistics*. New York: Harper & Row.

Tabachnick, B.G., & Fidell, L.S. (1989). *Using multivariate statistics* (2nd ed.). New York: Harper & Row.

Tabachnick, B.G., & Fidell, L.S. (1996). *Using multivariate statistics* (3rd ed.). New York: HarperCollins.

Thompson, B. (1986, November). *Two reasons why multivariate methods are usually vital*. Paper presented at the annual meeting of the Mid-South Educational Research Association, Memphis.

Thompson, B. (1990). MULTINOR: A FORTRAN program that assists in evaluating multivariate normality. *Educational and Psychological Measurement, 50,* 845-848.

Thompson, B. (1994, February). Why multivariate methods are usually vital in research: Some basic concepts. In *Paper presented as a Featured Speaker at the biennial meeting of the Southwestern Association for Research in Human Development, Austin, T. (ERIC Document Reproduction Service No. ED 367 687)*

Appendix A
SPSS Commands for Female Group (n=64)

```
SET BLANKS=SYSMIS UNDEFINED=WARN printback=list.
TITLE 'MULTINOR.SPS   tests multivar normality graphically****'.
COMMENT The original MULTINOR computer program was presented, with examples, in:
COMMENT  Thompson, B. (1990). MULTINOR: A FORTRAN program that assists in
COMMENT  evaluating multivariate normality. Educational and Psychological Measurement, 50, 845-848.
COMMENT  The data source for the example are from:  George, D. J., and Mallery, P. (1999). SPSS for
COMMENT  Windows step by step. Boston: Allyn & Bacon.
COMMENT  Here there are 3 variables for which multivariate normality is being confirmed.
DATA LIST
  FILE='a:normgrad.dat' FIXED RECORDS=1 TABLE
  /1 gender 1 ethnicit 3 year 5 lowup 7 section 9 prevgpa 11-14  (1) final 16-17 (1)  total 19-21 (1)  .
list variables=all/cases=9999/format=numbered .
COMMENT  'y' is a variable automatically created by the program, and does not have to modified for different data sets.
select if (gender eq 1) .
compute y=$casenum .
print formats y(F5) .
regression variables=y prevgpa to total/
  descriptive=mean stddev corr/
  dependent=y/enter prevgpa to total/
  save=mahal(mahal) .
sort cases by mahal(a) .
execute .
list variables=y prevgpa to total mahal/cases=9999/format=numbered .
COMMENT  In the next TWO lines, for a given data put the actual n in place of the number '64' used for the example data set.
loop #i=1 to 64 .
compute p=($casenum - .5) / 64.
COMMENT  In the next line, change '3' to whatever is the number of variables. The p critical value of
COMMENT chi square for a given case is set as [the case number (after sorting) - .5] / the sample size].
if (gender eq 1) chisq=idf.chisq(p,3) .
end loop .
print formats p chisq (F8.5) .
list variables=y p mahal chisq/cases=9999/format=numbered .
plot
  vertical='chi square'/
  horizontal='Mahalanobis distance'/
  plot=chisq with mahal .
```

Appendix B
SPSS Commands for Male Group

```
SET BLANKS=SYSMIS UNDEFINED=WARN printback=list.
TITLE 'MULTINOR.SPS   tests multivar normality graphically****'.
COMMENT ********************************************************.
COMMENT The original MULTINOR computer program was presented, with examples, in:
COMMENT     Thompson, B. (1990). MULTINOR: A FORTRAN program that assists
COMMENT      in evaluating multivariate normality. Educational and Psychological Measurement, 50, 845-848.
COMMENT     The data source for the example are from:
COMMENT     George, D. J., and Mallery, P. (1999). SPSS for Windows step by step. Boston: Allyn & Bacon.
COMMENT  Here there are 3 variables for which multivariate normality is being confirmed.
DATA LIST
  FILE='a:normgrad.dat' FIXED RECORDS=1 TABLE
             /1 gender 1 ethnicit 3 year 5 lowup 7 section 9 prevgpa 11-14  (1)    final 16-17 (1)
  total 19-21 (1)  .
list variables=all/cases=9999/format=numbered .
COMMENT  'y' is a variable automatically created by the program,
COMMENT  and does not have to modified for different data sets.
select if (gender eq 2) .
compute y=$casenum .
print formats y(F5) .
regression variables=y prevgpa to total/
  descriptive=mean stddev corr/
  dependent=y/enter prevgpa to total/
  save=mahal(mahal) .
sort cases by mahal(a) .
execute .
list variables=y prevgpa to total mahal/cases=9999/format=numbered .
COMMENT  In the next TWO lines, for a given data set put the
COMMENT  actual n in place of the number '41' used for the
COMMENT  example data set.
loop #i=1 to 41 .
compute p=($casenum - .5) / 41.
COMMENT  In the next line, change '3' to whatever is the number
COMMENT  of variables.
COMMENT       The p critical value of chi square for a given case
COMMENT  is set as [the case number (after sorting) - .5] / the
COMMENT  sample size].
if (gender eq 2) chisq=idf.chisq(p,3) .
end loop .
print formats p chisq (F8.5) .
list variables=y p mahal chisq/cases=9999/format=numbered .
plot
  vertical='chi square'/
  horizontal='Mahalanobis distance'/
  plot=chisq with mahal .
```

Appendix C
SPSS Syntax for Evaluating Univariate and Bivariate Normality

```
PPLOT
  /VARIABLES=prevgpa
  /NOLOG
  /NOSTANDARDIZE
  /TYPE=Q-Q
  /TIES=MEAN
  /DIST=NORMAL .
GRAPH
  /HISTOGRAM=prevgpa .
EXAMINE
  VARIABLES=prevgpa final total
  /PLOT BOXPLOT STEMLEAF HISTOGRAM NPPLOT
  /COMPARE GROUP
  /STATISTICS DESCRIPTIVES
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL .
GRAPH
  /SCATTERPLOT (BIVAR)=prevgpa WITH total
  /MISSING=LISTWISE .
PLOT
  /VERTICAL='prevgpa' REFERENCE (6,4)
  /HORIZONTAL='total' REFERENCE (6,7)
  /PLOT=prevgpa WITH total .
GRAPH
  /SCATTERPLOT (BIVAR)=prevgpa with final
  /MISSING=LISTWISE .
PLOT
  /VERTICAL='prevgpa' REFERENCE (6,4)
  /HORIZONTAL='final' REFERENCE (6,9)
  /PLOT=prevgpa WITH final .
GRAPH
  /SCATTERPLOT (BIVAR)=final with total
  /MISSING=LISTWISE .
PLOT
  /VERTICAL='final' REFERENCE (6,9)
  /HORIZONTAL='total' REFERENCE (6,7)
  /PLOT=final WITH total .
COMMENT  is set as [the case number (after sorting) - .5] / the
COMMENT  sample size].
compute p=($casenum - .5)/62 .
compute chisq=idf.chisq(p,3) .
end loop .
print formats p chisq (F8.5) .
list variables=y p mahal chisq/cases=9999/format=numbered .
plot
  vertical='chi square'/
  horizontal='Mahalanobis distance'/
  plot=chisq with mahal .
```

Appendix D
SPSS Commands for New Dependent Variable

```
SET BLANKS=SYSMIS UNDEFINED=WARN printback=list.
TITLE 'MULTINOR.SPS   tests multivar normality graphically****'.
COMMENT ********************************************************.
COMMENT   The original MULTINOR computer program was presented,
COMMENT   with examples, in: Thompson, B. (1990). MULTINOR: A FORTRAN program that assists
COMMENT    in evaluating multivariate normality. Educational and Psychological Measurement, 50, 845-848.
COMMENT     The data source for the example are from:
COMMENT     George, D. J., and Mallery, P. (1999). SPSS for Windows step by step. Boston: Allyn & Bacon.
COMMENT ********************************************************.
COMMENT  Here there are 3 variables for which multivariate normality is being confirmed.
DATA LIST
  FILE='a:norgrades.txt' FIXED RECORDS=1 TABLE
  /1 quiztot 1-2 (1) final 4-5 (1) total 7-9 (1) .
list variables=all/cases=9999/format=numbered .
COMMENT  'y' is a variable automatically created by the program,
COMMENT  and does not have to modified for different data sets .
compute y=$casenum .
execute .
print formats y(F5) .
regression variables=y quiztot to total/
  descriptive=mean stddev corr/
  dependent=y/enter quiztot to total/
  save=mahal (mahal) .
sort cases by mahal(a) .
execute .
list variables=y quiztot to total mahal/cases=9999/format=numbered .
COMMENT  In the next TWO lines, for a given data set put the
COMMENT  actual n in place of the number '62' used for the
COMMENT  example data set.
loop #i=1 to 62 .
COMMENT  In the next line, change '3' to whatever is the number of variables.
COMMENT       The p critical value of chi square for a given case
```

Using Partial Residual Plots in Assessing and Improving the Construct Validity of Multiple Regression Models

Cam-Loi Huynh, University of Manitoba

Advantages of *partial residual plots* over *residual plots* in regression analysis are discussed and illustrated by empirical examples. A variation of partial residual equation is introduced and an effective procedure to use this revised form in identifying the proper transformation for achieving linearity and variance stabilization is presented. Essentially, the transformed predictors are identified by partial residual plots and introduced into the regression model to improve the regression fit. Uses, limitations and strengths of partial residual plots are discussed.

In formulating the multiple regression model, researchers often feels strongly that an explanatory variable (x_j) included in the model influences the response (y). But, they are not sure whether it is the variable (x_j), as they happen to measure it, or some function $g(x_j,$ that is linearly related to the mean of the response. This is often because the jth regression coefficient is smaller than expected, statistically insignificant, or of the "wrong" sign. Unfortunately, estimates of partial regression coefficients and summary statistics such as R^2, F and t are unable to detect sources of the failure to yield good fit (For a good discussion on this aspect, see Belsley, Kuh & Welsch, 1980; Cook & Weisberg, 1982).

The standard recommendation in assessing model-data fit is to plot *residuals* (e) and *studentized residuals* (r) against the independent variables (e.g., Cook & Weisberg, 1982, Chapter 2; Draper and Smith, 1981, Chapter 3). These plots help the researcher in (i) detecting outliers, (ii) assessing the presence or absence of variance heterogeneity, and (iii) determining if a transformation of the explanatory variable is needed or if another term (e.g., a quadratic term) needs to be added. In addition to providing these information, *partial residual plots* enable the researcher in (iv) assessing the importance of x_j (in terms of predicting power for y) in the presence of all other independent variables and (v) evaluating the importance of nonlinearity among the of x_j variables and choosing the appropriate transformation more precisely (Larsen & McCleary, 1972).

In this paper, the comparative properties of residuals and partial residual plots are discussed and illustrated by an empirical example. A variation of partial residuals is introduced and an effective procedure to use this revised form for improving the fit of multiple regression models is presented and examined by means of simulation data. Finally, comments on the uses, limitations, and strengths of partial residual plots are given.

Empirical Properties of Residualsand Partial Residuals

In this paper, the lower-case letters x and y and upper-case letters $\boldsymbol{X}$ and $\boldsymbol{Y}$ are used to represent vectors and matrices of the independent and dependent variables, respectively.

Suppose a researcher considered the regression model

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon' = \boldsymbol{X_A \beta} + \varepsilon' \qquad (1)$$

(called the "restricted" model), where ε' represents the associated (but unknown) residual term, X_A is an (n by k + 1) design matrix of the intercept and independent variables, and $\boldsymbol{\beta}$ is a (k + 1) vector of regression coefficients. He subsequently added an independent variable x_q to improve its fit and interpretation of its parameters. As a result, the regression model

$$y = \boldsymbol{X_A \beta} + \beta_q x_q + \varepsilon \qquad (2)$$

(called the "observed" model), is obtained, where the residual term ε is estimated by e. Suppose the outcome was found unsatisfactory (e.g., insignificant increase in R^2, unexpected sign of β_q or some nonlinear relationship revealed in the plot of the predicted values $\hat{y}$ against x_q). Now, the researcher wants to determine the form $g(x_q)$ such that

$$y = \boldsymbol{X_A \beta} + \gamma g(x_q) + \varepsilon^* \qquad (3)$$

(called the "correct" model), where γ denotes the *qth* slope coefficient, would yield a substantially better fit than (2).

The computational formula for sample residuals in the fitted regression equation of

$$\hat{y} = b_0 + b_1 x_1 + \ldots + b_k x_k + b_q x_q, \qquad (2')$$

an estimate of the "observed" model (2), is expressed as

$$e = y - \hat{y}, \qquad (4)$$

and the associated *partial residuals* are defined as

$$r = e + b_q x_q \qquad (5)$$

The equation (5) was first discovered by Ezekiel (1924) and reintroduced by Larsen & McCleary (1972). Partial residuals are also called the "component-plus residuals" by Wood (1973).

Residual and partial residual plots are obtained when e and r are plotted against x_q, respectively. Besides these graphical methods, three other main diagnostic plots for explanatory variables are (i) internal and external *studentized residual plots* (Cook & Weisberg, 1982, pp. 18-20), (ii) *added variable plots* (Wood, 1973), (iii) *partial regression leverage plots* (first used by Draper & Smith, 1966, p. 112; reintroduced by Mosteller & Tukey, 1977, pp. 344-345

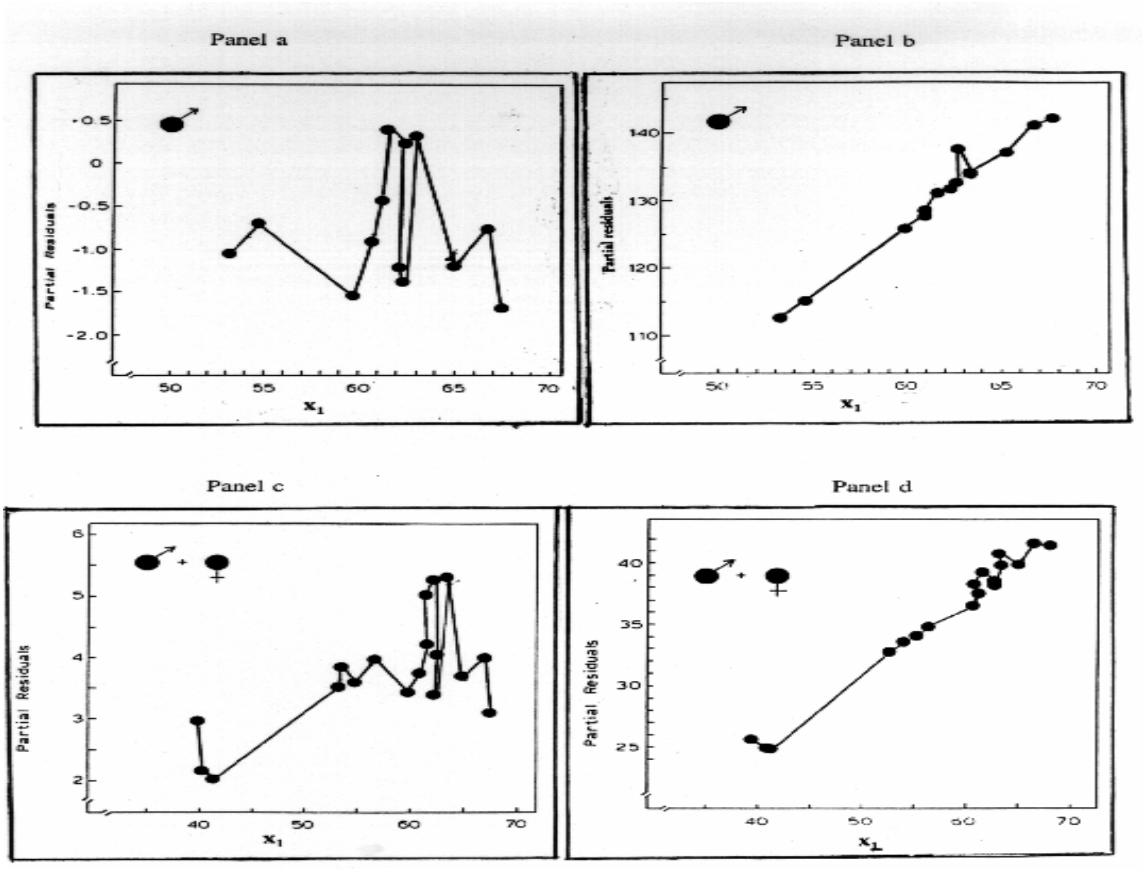


Figure 1. Plots of Residuals (e) and Partial Residuals (r) Against x_1 in Example 1. (Knwolton et al., 1980).

and 374-376; and advocated by Sall, 1990). Among these five methods, partial residual plots are easier to construct and simpler to understand than both added variable plots and partial regression leverage plots (Atkinson, 1985, p. 73). Moreover, the use of partial residual plots enables the researcher to determine more precise forms of nonlinear transformation than by using other plots (Gunst and Mason, 1980, Chapter 7).

If the relationship between x_q and y is linear, the plot of r against x_q should show data points distributed along a non-horizontal line. Moreover, its slope represents b_q in (5). On the contrary, if the relationship is nonlinear, the plot should indicate the nature of the transformation that is required to demonstrate a linear relationship. The following example serves to illustrate these properties.

Example 1. Knwolton et al. (1980) studied some physiological and performance characteristics of athletes in the sport of competitive orienteering. In particular, they used three variables (x_1 = maximal aerobic power, x_2 = years of experience, x_3 = anaerobic power and x_4 = blood lactate) to predict performance (y). For the males sample, the partial residual plots for x_2, x_3, and x_4 showed linear trends but the plot for x_1 was indicative of a quadratic relationship (Figure 1, Panel a). By introducing the variable x_5 as the square of x_1 into the regression equation, a nearly straight line was observed in the partial residual plot of r against x_1 (Figure 1, Panel b). The same findings were found for the total sample (males and females) before transformation of x_1 (Figure 1, Panel c), and after the introduction of x_1^2 (Figure 1, Panel d).

Theoretical Properties of Residuals and Partial Residuals

First, it will show that the plot of the residual (e) against x_q will not generally reveal the shape of the function $g(x_q)$. Next, the forms of partial residual (r) that can reveal $g(x_q)$ will be discussed. The sample residuals of the fitted model (2) can be rewritten as

$$e = (\boldsymbol{I} - \boldsymbol{H})y, \tag{6}$$

where $\boldsymbol{I}$ is the identity matrix of order n, $\boldsymbol{H}=\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}$ is an idempotent matrix, and $\boldsymbol{X} = (\boldsymbol{X}_A\ x_q)$ is an (n x q+1) augmented matrix (for $q = k + 1$). The expected value of the residual term is given as

$$\begin{aligned} E(e) &= (\boldsymbol{I} - \boldsymbol{H})(X\beta + \beta_q g(x_q) + \varepsilon^*) \\ &= (\boldsymbol{I} - \boldsymbol{H})(\beta_q g(x_q) \end{aligned} \tag{7}$$

because ($\boldsymbol{I}$ - $\boldsymbol{H}$) is orthogonal to both $X\beta$ and ε^*. It is

well known that the sample mean of e for regression models with the intercept is zero because

$$\Sigma e = 1'(\boldsymbol{I} - \boldsymbol{H})(\textit{estimate of } \beta_q g(x_q)) = 0, \qquad (8)$$

since $1'(\boldsymbol{I} - \boldsymbol{H}) = 0$ where 1 is a vector of unity. Notice that the mean of the residual estimate ($\bar{e}$) is zero regardless of the form of $g(x_q)$. For simplicity, let g denote $g(x_q)$. First, in the "ideal" case of $g = x_q$, the fitted model (2) is correct, $E(e) = 0$ and the residual plot would display a random pattern around 0. On the other hand, if g is any linear combination of the columns of $\boldsymbol{X_A}$ then $E(e)$ is also zero but the residual plot would disclose the form of $(\boldsymbol{I} - \boldsymbol{H})(\beta_q g)$, not of g. Finally, if g is a curvilinear function, say

$$g = \beta_{q0} + \beta_{q1}x_q + \beta_{q2}x_q^2,$$

then $E(e) = (\boldsymbol{I} - \boldsymbol{H})\beta_{q2}x_q^2$. The plot of e against x_q would indicate that the linearity assumption has been violated but the shape of the function g is still unknown since the residual plot only reveals the function $(\boldsymbol{I} - \boldsymbol{H})\beta_{q2}x_q^2$. Partial residual plots have been suggested as a more effective device than residual plots in detecting the function g (Larsen & McCleary, 1972 and Wood, 1973). Some theoretical properties of r can be explained by means of its expected value,

$$E(r) = E(e) + x_q E(b_q) = (\boldsymbol{I} - \boldsymbol{H})g + \varphi_q x_q, \qquad (9)$$

where $\varphi_q = E(b_q) = (\boldsymbol{D'D})^{-1}\boldsymbol{D'}g$, and $D = (\boldsymbol{I} - \boldsymbol{H}_A)x_q$, a residual obtained upon fitting x_q on the columns of $\boldsymbol{X}_A$, where $\boldsymbol{H_A} = \boldsymbol{X_A}(\boldsymbol{X_A'X_A})^{-1}\boldsymbol{X_A}$. If $\varphi_q = 0$, or $b_q \approx 0$, then $E(r) = E(e)$, implying that the "restricted" model (1) is tenable and both the residual and partial residual plots would indicate the insignificance of x_q in predicting y. On the other hand, if the "observed" model (2) is correct then $g = \beta_q x_q$ and $\varphi_q = \beta_q$ so that $(\boldsymbol{I} - \boldsymbol{H})g = 0$, $E(r) = 0$ and $E(r) = \beta_q x_q$. Then the partial residual plot against x_q would reveal a linear curve with its slope as an estimate of β_q. Finally, if g is not a linear combination of the columns of X_A, the plot of r versus x_q would indicate the nonlinear form of g (A proof of this effect has been shown by Manfield & Conerly, 1987).

A More Effective Partial Residual Form

From the preceding discussion, it becomes clear that the standard definition of the partial residual would only be indicative of the slope of the function g. Suppose that the function g is monotonically (non)linear. For a more complete information, it is suggested that the intercept term (c) be added so that the revised form the partial residual becomes

$$r^* = c + b_q x_q + e \qquad (10)$$

where the values of c are to be determined. If the "observed" model (2) is "correct" then the expected values of r^* would be

$$E(r^*) = c + \beta_q x_q \qquad (11)$$

which a straight line with intercept c and slope β_q. In all cases, including when g is a nonlinear function of x_q, the plot of r^* versus x_q would reveal estimates of both the intercept (c) and slope (β_q). It worth noting that, if the constant c represents the estimate of regression intercept in fitting model (2) then equation (8) can be rewritten as

$$r^* = \bar{y} + b_1(x_1 - \bar{x}_1) + \ldots + b_k(x_k - \bar{x}_k) + b_q(x_q - \bar{x}_q) + e, \qquad (12)$$

which is the same as equation 2.9 in Larsen and McCleary, 1972, p. 785 and equation 1.2 in Mallows, 1986, p. 313.

Before data collection, the researcher may not have any knowledge about the form of $g(x_q)$. When the transformed variable $w(x_q)$ is used as an estimate of $g(x_q)$, it results in what will be referred to as the "estimated" model

$$y = X_A\beta + \gamma w(x_q) + \varepsilon^*. \qquad (3')$$

Smith (1972) presented examples of linearizing regression equations, such as (3'), by manipulating the constant term c. Once a transformed model for y has been decided (say, $w = exp(y + c)$), it requires to plot only a few points of w with different values of c for the curve that is most linear to be identified. By applying Smith (1972)'s technique, it can be shown that r^* is superior to r for linearity and variance stabilization purposes. The illustration is quite easy for the two common forms of w, namely, logarithmic and power (or root) transformations. First, consider the logarithmic transformation $w = log_{10}(x_q + c)$, where $c > 0$, a vector of constant values to be determined. The effect of the logarithmic transformation w can be described better in terms of the inverse function $10^w = x_q + c$ for it implies that $x_q = - c + 10^w$. The graph of $x_q = - c + 10^w$, which represents an estimate of $g(x_q)$, varies continuously from an exponential curve when $c = 0$ to a linear line when c is large. As a strategy, one can fit model (3') $\hat{y} = X_A\beta + \gamma w(x_q) = X_A\beta + \gamma log_{10}(x_q + c)$ repeatedly by increasing the value of c until the improvement in R^2 becomes insignificant. Secondly, consider the power transformation $w = x_q^c$ for $-1 < c < 1$ with the associated inverse function of $x_q = w^{1/c}$. Its graph varies from a hyperbolic curve when $c = -1$ to an exponential curve as $c \approx 0$, and a line when $c = 1$. A strategy of repeated fitting of model (3') with positive values of c, but less than 1, can be applied. It is well-known that if variances of the columns of the design matrix X are increased proportionally to their means then one can use square root transformation on the columns for variance stabili- zation. On the other hand, if variances of the columns are proportional to the coefficients of variation (σ_j/μ_j, $j = 1, 2, \ldots, q$) then the logarithmic transformation may be used for variance stabilization (Drapper & Smith, 1981, pp. 146-148, 237-240). In all of these cases, if a constant c is added to the transformed functions, the accuracy of w as an

Table 1. Generated Data for Examples 2 and 3.

Example 2 (Logarithmic Transformation)					Example 3 (Power Transformation)				
y	x_1	x_2	x_q	g_x	y	x_1	x_2	x_q	g_x
2.236	-0.836	9.358	9.358	2.236	4.609	2.236	-0.836	0.836	1.128
1.381	-0.726	3.979	3.979	1.381	2.979	1.381	-0.726	0.662	1.318
0.058	-0.130	1.060	1.060	0.058	2.469	0.058	-0.130	0.595	1.416
0.946	1.237	2.575	2.575	0.946	4.223	0.946	1.237	0.986	1.009
-1.910	0.748	0.148	0.244	-1.410	1.453	-1.910	0.748	0.455	1.696
0.393	0.855	1.482	1.482	0.393	3.771	0.393	0.855	0.931	1.049
1.157	0.045	3.182	5.246	1.657	3.387	1.157	0.045	0.485	1.624
1.304	-0.092	3.685	3.685	1.304	3.654	1.304	-0.092	0.691	1.281
-0.039	-0.481	0.962	0.962	-0.039	1.935	-0.039	-0.481	0.910	1.065
0.570	0.596	1.767	2.914	1.070	5.232	0.570	0.596	0.178	3.176
-1.015	2.406	0.362	0.597	-0.515	3.569	-1.015	2.406	0.489	1.614
1.130	-0.530	3.095	5.102	1.630	3.462	1.130	-0.530	0.371	1.945
0.193	-0.793	1.213	1.213	0.193	3.924	0.193	-0.793	0.164	3.356
-0.953	1.159	0.385	0.385	-0.953	3.799	-0.953	1.159	0.302	2.232
-0.225	1.077	0.799	1.317	0.275	5.424	-0.225	1.077	0.118	4.176
1.289	-0.345	3.631	3.631	1.289	3.121	1.289	-0.345	0.793	1.169
0.991	0.717	2.693	2.693	0.991	4.198	0.991	0.717	0.598	1.411
-0.840	2.042	0.432	0.711	-0.340	5.995	-0.840	2.042	0.105	4.531
0.055	-1.304	1.057	1.742	0.555	0.678	0.055	-1.304	0.846	1.118
-0.379	-0.847	0.684	1.128	0.121	0.875	-0.379	-0.847	0.800	1.161
-0.320	-1.085	0.726	0.726	-0.320	3.482	-0.320	-1.085	0.237	2.623
1.541	-1.706	4.670	7.700	2.041	1.640	1.541	-1.706	0.856	1.110
0.639	0.232	1.895	3.125	1.139	2.554	0.639	0.232	0.821	1.141
-0.989	1.328	0.372	0.613	-0.489	6.158	-0.989	1.328	0.084	5.251
-0.423	0.165	0.655	0.655	-0.423	5.908	-0.423	0.165	0.003	5.167
1.801	-0.179	6.056	6.056	1.801	5.236	1.801	-0.179	0.810	1.152
-0.274	1.360	0.760	1.254	0.226	3.609	-0.274	1.360	0.400	1.848
1.948	-0.466	7.016	7.016	1.948	8.286	1.948	-0.466	0.074	5.721
1.046	0.487	2.846	4.692	1.546	9.536	1.046	0.487	0.054	7.069
1.371	0.877	3.938	3.938	1.371	4.415	1.371	0.877	0.940	1.042

estimator of $g(x_q)$ can be improved by determining the required values of c. As demonstrated in the following examples, the determination of c can be achieved after a few trial-and-error attempts.

Example 2. The random variables y and x_1 were generated as standard normal whereas x_2 and x_q as exponential, namely $x_2 = e^Y$ and $x_q = e^{y + .5}$, respectively. Their values are reported in Table 1. The transformed variable (w) was obtained according to the function $w = log_{10}(x_q + c)$. The resulting regression equations and corresponding R^2 are listed in Table 2. The largest value of R^2 corresponds to $c = 0$ so that $w = log_{10}(x_q)$. The improvement in R^2 due to the addition of x_q and then replacing it by w can be tested by the method of comparing two nonnested multiple regression models (Graybill & Iyer, 1994, pp. 309-313). In nonnested regression models, there are predictors in one model that are not found in the other model and there may be some predictors that occur in both. The test of the null hypothesis H_0: $R^2_A = R^2_B$ is the same as the test of H_0: $\sigma^2_A = \sigma^2_B$, where σ^2_A and σ^2_B are the sum of square of errors (*SSE*) in the two regression models A and B, respectively. If the 100(1 - α)% confidence of σ_B/σ_A contains the value of 1 then the null hypothesis is retained. On the other hand, if the upper bound of the confidence interval is less than 1 then the null hypothesis is rejected in supporting the alternative hypothesis that $\sigma_B < \sigma_A$, or $R^2_B > R^2_A$, which in turn implies that model B is better than model A. Similar arguments applies, but in favor of model A if the lower bound of the confidence interval is larger than 1. The confidence intervals reported in Table 3 indicates that the "estimated" model is statistically superior to both the "restricted" and "observed" models.

Example 3. A regression model similar to the one considered in Cook and Weisberg (1994) is studied. The variable y was generated according to the function

$$y = 1 + x_1 + x_2 + x_q^{-0.67} + \varepsilon, \qquad (13)$$

where x_1 and x_2 were normally distributed, ε was an independent normal with mean 0 and variance .025, and x_q was an uniform random variable. Their derived values are reported in Table 1. The transformed

Table 2. Effects of Changing c in Logarithmic and Power Transformations

Example 2	Logarithmic Transformation
Model 1 (restricted)	$\hat{y} = -0.476 - 0.129x_1 + 0.388x_2$, $R^2 = .800$
Model 2 (observed)	$\hat{y} = -0.564 - 0.096x_1 + 0.221x_2 + 0.168x_q$, $R^2 = .816$
Model 3 (estimated)	$w = Log(x_q + c)$
$c = 0.0$	$\hat{y} = -0.344 - 0.042x_1 + 0.101x_2 + 0.815g_x$, $R^2 = .955$
$c = 0.5$	$\hat{y} = -0.825 - 0.051x_1 + 0.075x_2 + 1.121g_x$, $R^2 = .932$
$c = 1.0$	$\hat{y} = -1.310 - 0.056x_1 + 0.069x_2 + 1.348g_x$, $R^2 = .917$
$c = 1.5$	$\hat{y} = -1.790 - 0.059x_1 + 0.069x_2 + 1.542g_x$, $R^2 = .906$
$c = 5.0$	$\hat{y} = -4.952 - 0.069x_1 + 0.094x_2 + 2.556g_x$, $R^2 = .866$
Example 3	**Power Transformation**
Model 1 (restricted)	$\hat{y} = 5.874 - 0.802x_1 + 0.261x_2$, $R^2 = .011$
Model 2 (observed)	$\hat{y} = 12.038 + 0.465x_1 - 0.251x_2 + 12.696x_q$, $R^2 = .176$
Model 3 (estimated)	$w = x_q^c$
$c = -2.00$	$\hat{y} = 3.199 + 1.046x_1 + 1.207x_2 + 0.0004g_x$, $R^2 = .974$
$c = -0.67$	$\hat{y} = 1.048 + 1.087x_1 + 0.841x_2 + 0.999g_x$, $R^2 = .998$
$c = -0.30$	$\hat{y} = -9.923 + 1.183x_1 + 0.238x_2 + 9.902g_x$, $R^2 = .923$
$c = 0.30$	$\hat{y} = 28.237 + 0.858x_1 - 0.453x_2 - 29.896g_x$, $R^2 = .426$
$c = 0.70$	$\hat{y} = 14.971 + 0.587x_1 - 0.351x_2 - 15.977g_x$, $R^2 = .240$

Table 3. Steps in Computing the Two-Sided Confidence Intervals for σ_B/σ_A using the Bonferroni Method ($\alpha = 0.05$).

Step	Example 2 (Logarithmic Transformation)
(1)	The 97.5% 2-sided confidence interval for σ_A in the "observed"model ($L_A = .354$, $U_A = .669$) where $L_A = \{SSE(A)/\chi^2_{1-\alpha/4;n-3-1}\}^{0.5} = \{5.624/44.762\}^{0.5}$, and $U_A = \{SSE(A)/\chi^2_{\alpha/4;n-3-1}\}^{0.5} = \{5.624/12.567\}^{0.5}$
(2)	The 97.5% 2-sided confidence interval for σ_B in the "estimated" ($L_B = .176$, $U_B = .332$) where $L_B = \{SSE(B)/\chi^2_{1-\alpha/4;n-3-1}\}^{0.5} = \{.385/44.762\}^{0.5}$ and $U_B = \{SSE(B)/\chi^2_{\alpha/4;n-3-1}\}^{0.5} = \{1.385/12.567\}^{0.5}$
(3)	The 95% 2-sided confidence interval for σ_B/σ_A: ($L_B/U_A = .263$, $U_B/L_A = .938$)
Step	**Example 3 (Power Transformation)**
(1)	The 97.5% 2-sided confidence interval for σ_A in the "observed"model ($L_A = 6.817$, $U_A = 12.866$) where $L_A = \{SSE(A)/\chi^2_{1-\alpha/4;n-3-1}\}^{0.5} = \{2080.255/44.762\}^{0.5}$, and $U_A = \{SSE(A)/\chi^2_{\alpha/4;n-3-1}\}^{0.5} = \{2080.255/12.567\}^{0.5}$
(2)	The 97.5% 2-sided confidence interval for σ_B in the "estimated" ($L_B = .367$, $U_B = .693$) where $L_B = \{SSE(B)/\chi^2_{1-\alpha/4;n-3-1}\}^{0.5} = \{6.028/44.762\}^{0.5}$ and $U_B = \{SSE(B)/\chi^2_{\alpha/4;n-3-1}\}^{0.5} = \{6.028/12.567\}^{0.5}$
(3)	The 95% 2-sided confidence interval for σ_B/σ_A: ($L_B/U_A = .029$, $U_B/L_A = .102$)

variable was determined to be $w = x_q^c$. As expected, the largest value of R^2 was found associated with c = -0.67. The improvement in R^2 for the three regression models, "restricted" (x_1, x_2), "observed" (x_1, x_2, x_q) and "estimated" (x_1, x_2, w), were tested by the method of comparing two nonnested multiple regression models. The resulting confidence intervals in Table 3 indicate that the "estimated" model is statistically superior to the other models.

Procedure to Detect the Function $g(x_q)$

As a first step, the transform variable w (of the function $g(x_q)$) can be determined by examining the plots of residuals (e) and partial residuals (r, in equation 6) against x_q. Next, x_q is substituted by w in computing a fit for the initial "estimated" model. The significance of the improvement in R^2 can be assessed by the method of comparing two nonnested multiple regression models. The formula for w may be modified by examining the plots of expected residuals ($E(e)$) and partial residuals ($E(r^*)$) against x_q. The computational formulas for these expected values are given by

$$E(e) = [I - x_q(x_q'x_q)^{-1}x_q']g, \qquad (14)$$
$$E(r^*) = c + E(e) + \varphi_q x_q,$$

where $\varphi_q x_q = x_q(W'W)^{-1}W'g = x_q(x_q'x_q)^{-1}x_q'g$ so that

$$E(r^*) = c + [I - x_q(x_q'x_q)^{-1}x_q']x_q(x_q'x_q)^{-1}x_q'g + x_q(x_q'x_q)^{-1}x_q'g = c + g - x_qg + x_qg = c + g. \qquad (15)$$

The modification of w would be continued until the plot of $E(r^*)$ shows a nearly linear curve with a slope steeper (positively or negatively) than that in the plot of $E(e)$. The implementation of this procedure for the two preceding examples are studied below.

Example **2.** The reason why the sample mean of residuals are equal to 0 can be seen from the fact that values of e are balanced out on both sides of zero (Figure 2, Panel a). On the other hand, values of partial residuals (r) clearly indicate a steady positive trend as x_q increases (Figure 2, Panel b). It implies that x_q has a positively exponential distribution and the required transformation for linearizing its values would be a logarithmic function. The next step is to regress y on

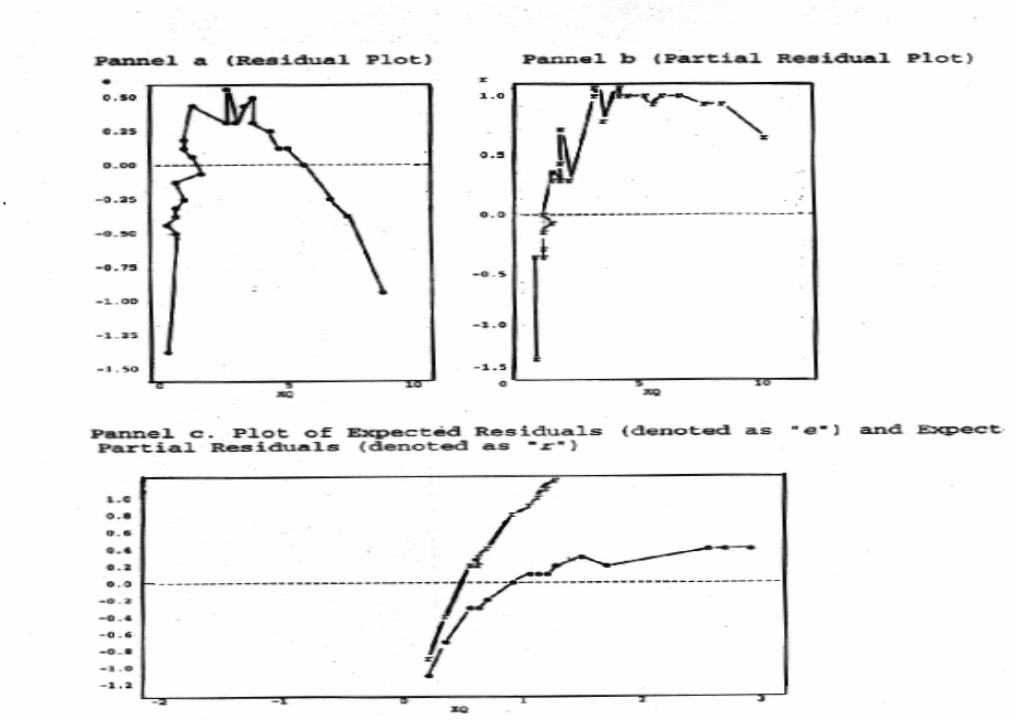


Figure 2. Plots of Residuals (e) and Partial Residuals (r*) and Their Expected Values ($E(e)$ and $E(r^*)$) Against x_a in Example 2

x_1, x_2 and w, where $w = log_{10}(x_q + c)$, with different values of c. By comparing the resulting R^2 and/or conducting the test of two nonnested regression models, one will know (at least statistically) if the "estimated" model is an improvement over the "observed" model. But how do we know that the model with the largest R^2 among those fitted would be acceptable or "correct" given the fact that R^2 can increase with an entry of even irrelevant independent variable into the regression model? The answer is found by observing the plots of the expected values $E(e)$ and $E(r^*)$ against x_q (Figure 2, Panel c) for the chosen "estimated" model. Whereas the plot of $E(e)$ reflects the nonlinearity nature of x_q, that of $E(r^*)$ shows a linear line with relatively steeper slope representing the strength of w in predicting y. In short, when the transformation is correctly determined, the resulting regression model would render largest R^2 and a graph of $E(r^*)$ with steepest-sloped curve.

Example 3. Even though the sample mean of residuals are equal to 0, this fact does not lend support to the tenability of the assumption of random error in the "observed" model. In Figure 3, Panel a, although most residual values lie below zero, they are cancelled out by the existence of a very large residual outlier. On the contrary, the partial residuals (r) clearly indicate a monotonic downward trend as x_q increases (Figure 3, Panel b). Therefore, $g(x_q)$ is deemed a negatively-sloped function so that the required transformation would be an inverted function (or negative root) of the form $w = x_q^c$, where $c < 0$. By comparing the resulting R^2 and conducting the test of two nonnested regression models, a satisfactory model can be identified with c . -.70. In this case, even if the true model is unknown, we still know that the model with the largest R^2 among those fitted would be "correct." This is because the plot of the expected values $E(r^*)$ against x_q shows an approximately linear line whereas the plot of $E(e)$ against x_q is clearly nonlinear (Figure 3, Panel c).

Discussion

Two uses of partial residual plots have been shown in the three examples discussed above. In Example 1, the objective is to improve the regression fit by introducing x_q^2 as an additional predictor of y. This strategy is applied mainly for meeting statistical assumptions of regression models (In this case, the linear relationship between y and its predictors). In Examples 2 and 3, the construct validity of x_q in predicting y can be improved by identifying w, an operational definition of the unknown function $g(x_q)$. The improvement in the resulting model serves not only to satisfy statistical assumptions but also to facilitate the model interpretation. This can be explained further from the fact that, even when all statistical assumptions are deemed satisfactory, multiple regression models still have construct validity problems (Winne, 1983). Huynh (2000) indicated that the effects of regressors in multiple regression models do not represent those of the constructs described by the original data since partial regression coefficients are actually computed for the residualized scores

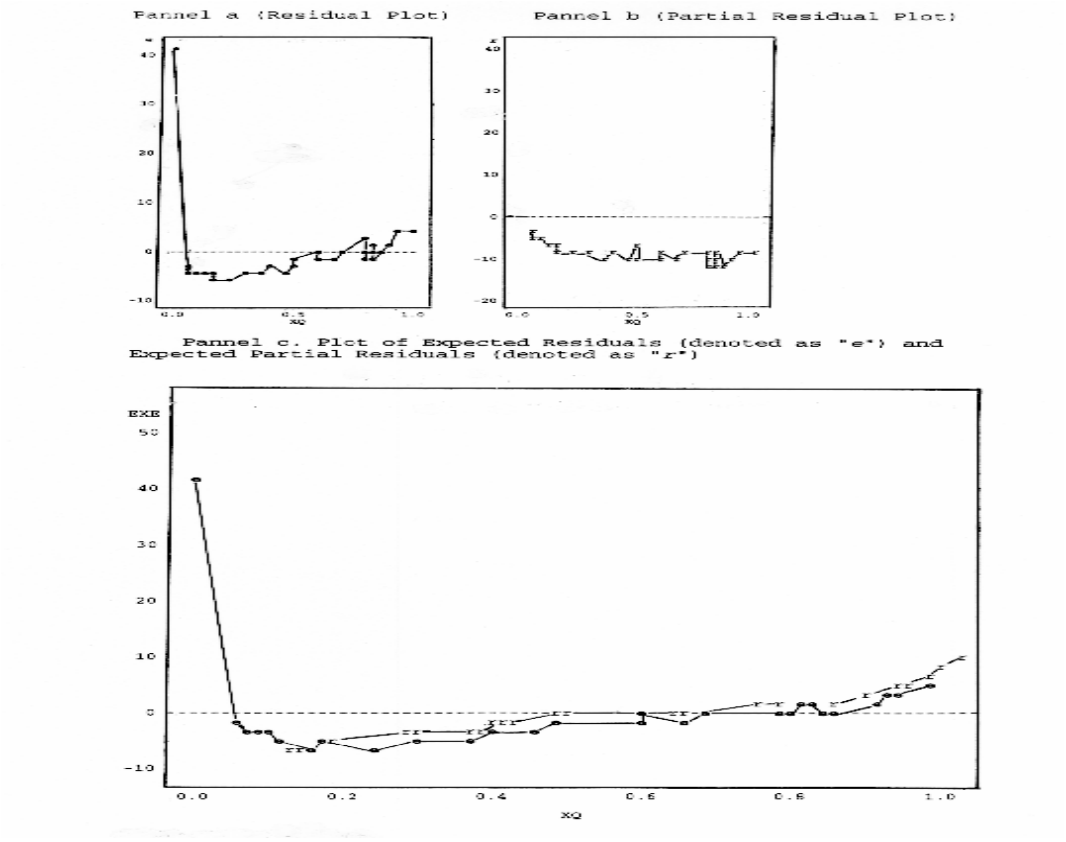


Figure 3. Plots of Residuals (*e*) and Partial Residuals (*r*) and Their Expected Values ($E(e)$ and $E(r^*)$) Against x_q in Example 3.

instead. The residualized score of the jth predictor (x_j) represents the residual term when x_j is regressed on the remaining predictors in the original multiple regression model. Therefore, in place of x_j, the relevant question is how the construct $g(x_j)$ can be reintroduced into the multiple regression. The procedure of examining partial residuals would be helpful for this purpose.

References

Atkinson, A. C. (1985). *Plots, transformations, and regression*. Oxford, UK: Clarendon Press.

Belsley, D. A., Kuh, E. & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York: Wiley.

Cook. R. D. & Weisberg, S. (1982). *Residuals and influence in regression*. London: Chapman & Hall.

Cook, R. D., & Weisberg, S. (1994). *An introduction to regression graphics*. New York: Wiley.

Cook, R. D. (1993). Exploring partial residual plots. *Technometrics,* 35(4), 351-362.

Draper, N., & Smith, H. (1981). *Applied regression analysis* (2nd ed.). New York: Wiley.

Ezekiel, M. (1924). A method of handling curvillinear correlation for any number of variables. *Journal of the American Statistical Association,* 19, 431-453.

Gunst, R. F., & Mason, R. L. (1980). *Regression analysis and its application*. New York: Chapman & Hall.

Knwolton, R.G., Ackerman,K.J., Fitzgerald, P. I., Wilde, S. W. & Tahamont, M. V. (1980). Physiological and performance characteristics of United States championship class orienteers. *Medicine and Science in Sports and Exercise,*12, 164-169.

Huynh, C-L. (2000). Extraneous variables and the interpretation of multiple regression coefficients. *Multiple Regression Viewpoints*. 26(1), 28-35.

Larsen, W. A., & McCleary, S. J. (1972). The use of partial residual plots in regression analysis. *Technometrics*, 14(3), 781-790.

Mallows, C. L. (1986). Augmented partial residual plots. *Technometrics,* 14, 313-320.

Mansfield, E. R. & Conerly, M. D. (1987). Diagnostic value of residual and partial residual plots. *The American Statistician,* 41(2), 107-116.

Sall, J. (1990). Leverage plots for general linear hypotheses. *American Statistician*, 44(4), 308-315.

Weisberg, S. (1985). *Applied linear regression* (2nd ed.). New York: Wiley.

Winne, P. H. (1983). Distortions of construct validity in multiple regression analysis. *Canadian Journal of Behavioral Science*, 15(3), 187-202.

Wood, F. S. (1973). The use of individual effects and residuals in fitting equations to data. *Technometrics,* 15, 677-695.

Rasch Measurement Instead of Regression

Benjamin D. Wright, University of Chicago
Kyle Perkins, Southern Illinois University
J. Kevin Dorsey, Southern Illinois University

This paper illustrates the use of Rasch measurement as an alternative to regression analysis to identify gout and non-gout patients who do and do not fit a gout variable constructed by the Rasch model. Unlike a numerical regression coefficient report, the clinician benefits from the one clear picture of the relationship among laboratory values and gout diagnosis which the Rasch model constructs.

The purpose of this paper is to demonstrate how a Rasch measurement model can be used to identify and describe the relationships among laboratory abnormalities among patients who have and have not been diagnosed with gout. Rasch analysis constructs a mathematical model of a gout variable, which identifies blood chemistries which co-occur and measures their utility as gout predictors. The model identifies patients whose pattern of blood chemistries do and do not fit the model. The purpose of this paper is to show how a blood chemistry profile can be organized to make a patient's diagnosis specific blood status immediately apparent to the clinician.

The Measurement Model

Rasch measurement construction applies a stochastic Guttman model to convert dichotomous, interval, and rating scale observations into linear measures (to which linear statistics can be usefully applied) and tests for goodness-of-fit to validate its item calibrations and patient measures. The basic fit statistic is a ratio of observed residual variance to expected residual variance and is near 1.00 when observed variance is comparable to expected.

In this application the Rasch model combines calibrations of blood chemistry items additively to patient measures to define observation probabilities. This stochastic conjoint additivity specifies a Guttman scale of probabilities to which the data are fit stochastically.

Rasch measurement estimates each parameter independently of all other parameters because its sufficient statistics exhaust all information in the data, which is associated with each parameter.

The unidimensional linear continuum to which the Rasch model fits the data provides estimates of item calibrations and patient measures which are not affected by the distributions of items or patients. Patient measures are test-free because their estimates are adjusted for the difficulty distribution of the items reported for that patient. All estimates are expressed in linear measures on a common scale defined by a single latent variable (Wright, 1999; Wright & Stone, 1971).

The Variables

Risk factors for gout have been studied intensively. The risk factor items used in this study are: uric acid, gender, age (at gout diagnosis), the presence or absence of diabetes, hypertension, kidney stones and diuretics, weight, height, body surface area, uric acid, cholesterol, triglycerides, urea nitrogen and creatinine.

Gout is a heterogeneous group of genetic and acquired diseases characterized by the deposition of monosodium urate monohydrate crystals in a joint. Alcohol, surgery, or trauma can trigger gout (Wolfe, 1991). Gout is chiefly a disease of men. Peak incidence occurs between ages 30 and 50 (Harris et al., 1999).

Further medical information can be found in Acheson et al., 1966; Berger & Yu, 1975; Campbell, 1988; Culleton et al., 1999; Evans et al., 1968; Garrick et al., 1972; Gibson & Grahame, 1974; Glynn et al., 1983; Murphy et al., 1982; Roubenoff, 1990; Roubenoff et al., 1991; Wolfe, 1991; Wolfe & Cathy, 1991; and Wyngaarden, 1988.

Method

Patient Selection

The computer records of a multi-specialty group practice were searched for patients with a gout diagnosis who had an office visit during a nine-month period. Of 91 charts available for review, 48 patients had information for all items under investigation.

Forty-eight patients without gout who had multi-channel chemistry profiles during a previous three-month period were matched pairwise by gender and age to the 48 gout patients.

Chart Review

At the first attack of gout, patients' gender, age, height, weight, urea nitrogen, creatinine, blood pressure, treatment with diuretics, and presence of insulin or non-insulin dependent diabetes mellitus were recorded. Kidney Stones were considered present if documented at any time in the patient's chart. Uric acid values were obtained while the patient was asymptomatic and not receiving allopurinol or uricosuric therapy. Cholesterol and triglyceride values were obtained after an overnight fast. Ninety-six observations were submitted for analysis: forty-eight gout and forty-eight non-gout patients, each observation having values recorded for the previously mentioned items.

Uniform Data Coding

For blood chemistries in mg/dl, height in inches and weight in pounds, each physical science metric value X was recoded linearly to nearest integer codes:

$$Y=9(X-MIN)/(MAX-MIN)$$

This coding simplifies the physical science metrics to 10 equal size steps labeled 0 through 9. The resulting codes are co-linear with the original physical science variables. (Table 2 gives the codes for uric acid, urea nitrogen and creatinine.)

Analysis

The data are analyzed by the WINSTEPS Rasch Analysis computer program (Linacre & Wright, 2000). WINSTEPS examines the complete data set, calculates fit statistics for each diagnostic item, uses a component analysis of data residuals to identify significant relationships among the diagnostic items and deletes items which do not contribute useful information. The result is the best linear variable for predicting gout, which these data support.

Results

Fifteen items of medical record information were provided for 96 patients. Forty-eight of the patients are a typical sample of patients diagnosed to have gout. The other 48 patients were selected so that each gout patient was matched by another patient similar in age and in gender but without a gout diagnosis.

Since our purpose is to explore the utility of a new way to analyze and display these kinds of data, we set aside prior expectations as to which information is supposed to predict gout. Instead, we begin our analysis with an open mind to find out how well this new method of analysis, implemented by WINSTEPS, can discover the best ways to predict gout from these data without cueing as to which patients are supposed to have gout and then to display this prediction in a graphical way that is clinically useful.

Unlike the regression approach, we do not use the presence/absence of a gout diagnosis as a "dependent variable" by which to narrow the combined effects of other, "independent variables". WINSTEPS can do this by anchoring patients on their gout diagnosis. But, for this article, we show instead what WINSTEPS can discover without being cued to detecting gout as its only object. We use WINSTEPS to look for the most general combination of the available medical record information, which maximizes a single measurement separation of these 96 patients, independent of their gout diagnosis.

We begin with all 15 original medical information items and step-by-step set aside items, which WINSTEPS misfit analysis shows are inconsistent with the construction of a single measure. The final steps in this process are reported in Table 1.

Table 1. Successive Deletions of Most Misfitting Blood Chemistry Item

INFIT MNSQ	INFIT ZSTD	MISFIT OUTFIT MNSQ	ORDER OUTFIT ZSTD	ITEM
		STEP	ONE	
1.29	1.90	*1.31	*2.0	Cholesterol
1.15	0.70	0.89	-0.5	Triglycerides
0.90	-0.50	0.95	-0.2	Urea Nitro
0.79	-0.90	0.79	-0.9	Creatinine
0.77	-1.80	0.78	-1.7	Uric Acid
		STEP	TWO	
1.47	2.10	*1.41	*1.30	Triglycerides
0.81	-0.80	1.06	0.20	Creatinine
0.84	-0.80	0.89	-0.60	Urea Nitro
0.75	-1.90	0.69	-2.40	Uric Acid

Note. * indicates deleted variables.

At each step we examine an item component factor analysis of data residuals to monitor dimensionality. By the time we have reduced our number of items from 15 to 11, it becomes obvious that surface, weight and gender imply a different measure for these patients than the four remaining blood chemistry items.

Figure 1 [WINSTEPS Table 23.2], "Finding the Variables from Rasch Residual Principal Components", shows the results of a principal component factor analysis of data residuals from the best single measure the 11 remaining items can support. The plot of item factor loadings against item measures shows a clear separation of male corpulence, clustered at the top of Figure 1, from the blood chemistries, clustered at the bottom. The location of a gout diagnosis in the center of the blood chemistry cluster at the bottom of Figure 1 implies that blood chemistry may provide better information about gout and also hypertension and a diuretic regimen than male corpulence among these 96 patients.

We could develop two measures of "gout", one based on male corpulence and another based on blood chemistry. This article is about the four blood chemistries appearing at the bottom of Figure 1.

During the step-wise analysis of the five blood chemistries shown in Table 1, WINSTEPS reports that the separation of patients by measure is improved by setting aside cholesterol and triglyceride information. After triglyceride is removed from the measurement model, Creatinine emerges as the next least informative blood chemistry item. We could set

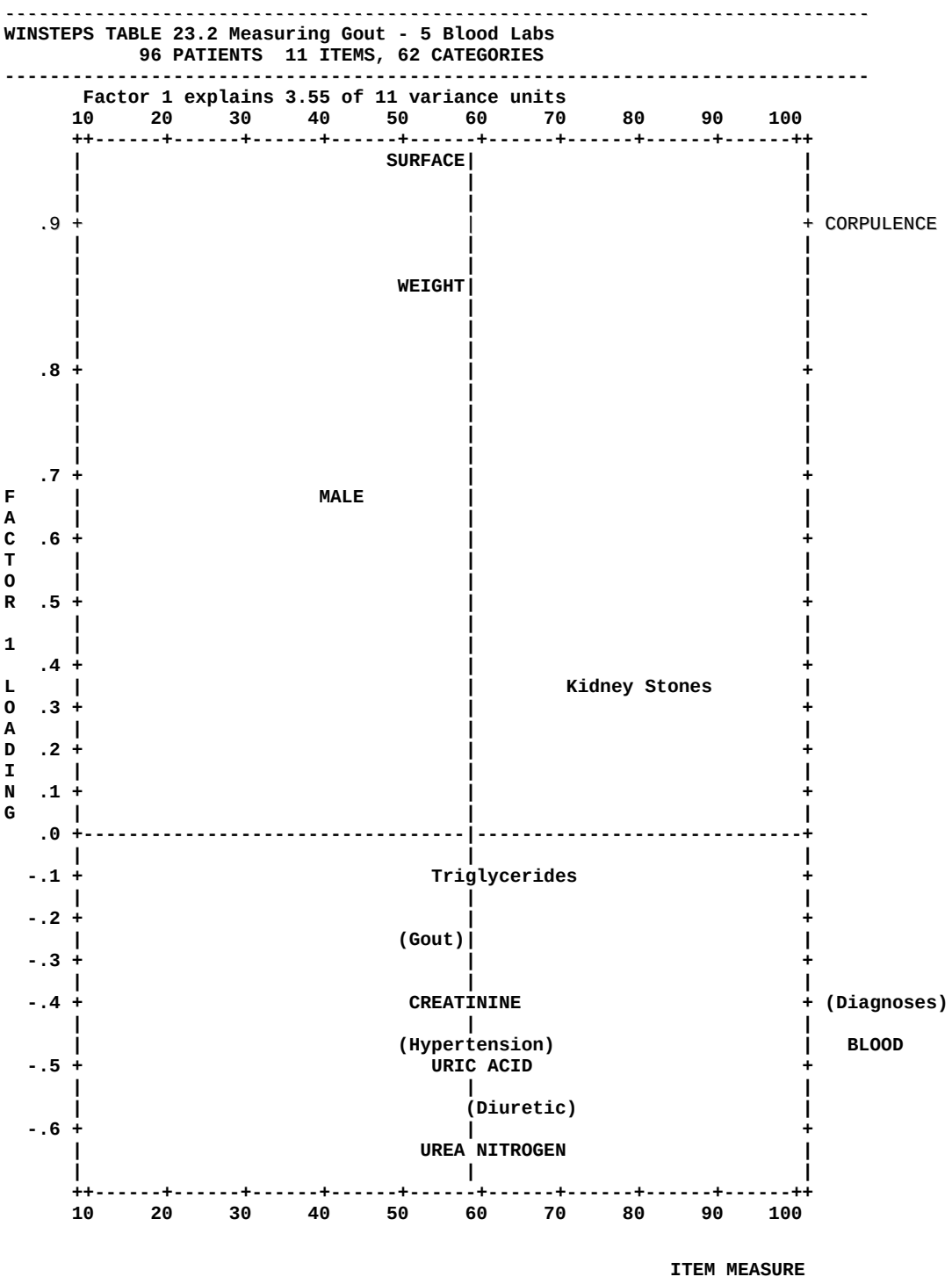


Figure 1. Finding the Variables from Rasch Residual Principal Components

```
-------------------------------------------------------------------------
WINSTEPS TABLE 2.2 Measuring Gout from 3 Blood Tests
      96 PATIENTS, 4 ITEMS, 32 UNIFORM CATEGORIES
-------------------------------------------------------------------------

EXPECTED SCORE: MEAN  (":" INDICATES HALF-SCORE POINT)
10   20    30    40    50    60    70    80    90   100
|-----+-----+-----+-----+-----+-----+-----+-----+-----|    BLOOD TEST
0   1 :    2       :     3    :(4):5:6: 7: 8  :   9   9   Urea Nitrogen
|                                                         |
0    1   :   2   :  3  :  4 :(5) 6  :  7  : 8  : 9    9   Uric Acid
|                                                         |
|                                                         |
0    0    :        1     :   (2) : 3:45:67:8  : 9     9   Creatinine

0                   NO    :   (YES)                     1     GOUT?

|-----+-----+-----+-----+-----+-----+-----+-----+-----|
10   20    30    40    50    60    70    80    90   100

    1           1     3  5 65 87 31121 21 1        PATIENTS WITH GOUT
                 T     S     M      S     T

                1
       2  3   2  0   5 9  8 51 11 1                PATIENTS WITHOUT GOUT
         T     S     M      S     T
```

Figure 2. The Complete Story on One Page

aside creatinine and concentrate on uric acid and blood urea nitrogen. But analysis with and without creatinine produces statistically equivalent results, and creatinine results dramatize an important finding concerning the non-linear relation between chemical measures and their medical implications. We establish a new medical variable defined by the observed relationships among the three blood chemistries: uric acid, urea nitrogen and creatinine.

Figure 2 [WINSTEPS Table 2.2], "The Complete Story on One Page", shows the resulting definition and patient discrimination of a gout diagnosis. This definition of a blood chemistry gout variable based on laboratory measures of uric acid, urea nitrogen and creatinine provides a compelling patient chart for the clinician.

Figure 2 puts the diagnosis of gout from patient chemical values of these three blood chemistries onto a simple, easy to read chart which lays out all of the blood chemistry information for this medical variable and also shows where each of the 96 patients measure on this variable. This chart makes visible in complete context the relation between the diagnosis of gout (the dependent variable in a regression analysis) and the predictive efficacy of the three blood chemistry variable (the independent variables in the regression analysis). The differences between this analysis and regression is that all results are visualizable on a common linear metric and no results are contaminated by missing data or sample dependent covariance.

The three top rows of 0-9 integers in Figure 2 mark out the medical measure positions of 10 equally spaced mg/dl chemical levels as they were coded uniformly into integers 0-9. Table 2 lists the mg/dl values linearly represented by each of these 0-9 codes. The vertical alignments of the codes in Figure 2 mark the mg/dl values of the three blood chemistries which match in their relative strength of "gout" implication. The integer codes for mg/dl values which the 1997 Merck Manual specifies as "too high" are in parentheses.

The fourth row of Figure 2 marks the predictor positions on this medical measure of the observed "Gout?" diagnosis: NO or (YES). The colon between NO and (YES) at a blood chemistry measure of 53 marks the point at which the estimated odds for "Gout?" are an even, 1 to 1. Estimated gout odds can be calculated for any measure position from 10 to 100 because on this scale each increment of 9 units triples the odds that a patient has gout. For example, since the estimated odds at blood chemistry measure 53 are even, the estimated odds become 3 to 1 at 53+9=62 and 9 to 1 at measure 71. In the other direction, the estimated odds for gout drop to 1 to 3 at 53-9=44 and 1 to 9 at 35.

Table 2. Variations in Medical Implications for Equal mg/dl Increases in Creatinine

Uniform Coding of Blood Chemistry Levels in mg/dl

Uniform Code	Uric Acid mg/dl	Urea Nitrogen mg/dl	Creatinine Mg/dl
0	2.10	00	0.7
1	3.30	05	1.0
2	4.50	10	1.3
3	5.70	15	1.6
4	6.90	20	1.9
5	8.10	25	2.2
6	9.30	30	2.5
7	10.50	35	2.8
8	11.70	40	3.1
9	12.90	45	3.4

Medical Measure Changes Implied by Equal mg/dl Increments of Creatinine

mg/dl Change	Code Change	MedMeasure Change	Mg/dl per MedMeas Unit
0.3	0-1	22	0.0136
0.3	1-2	18	0.0167
0.3	2-3	8	0.0375
0.3	3-4	3	0.1000
0.3	4-5	1	0.3000

In Figure 2 the horizontal spacing of all reference points and measures is uniformly linear in units of medical importance. (The uneven spacing of codes 0 to 9 in Figure 2 shows that these medical implications units are not collinear with the original chemical mg/dl units). This medical spacing enables rapid visual evaluation of the medical distance of any patient measure to the left or right of the colon at blood chemistry measure 53 to be sufficiently accurate for clinical purposes and even faster and less error prone than juggling odds.

When a patient's blood chemistries measure them below the "NO" at 44, we can advise them with some confidence that their blood chemistry does not imply gout. When, on the other hand, their measure exceeds the "(YES)" at 62, then our advice would have to be otherwise. We can show them their own position on the "Gout?" blood chemistry chart so that they can see for themselves where they stand with respect to a blood chemistry diagnosis of gout.

Because the WINSTEPS chart in Figure 2 maps the medical implications of the relationship between blood chemistry and "Gout?" probability in easy-to-read equal spacing, clinicians can find it easy to discover in their own practice where the best turning points are for the decisions their practice teaches them to make.

The "Gout?" diagnosis row serves the same purpose as gout predictions derived from a regression analysis. In this application, however, the prediction is no longer twisted by the incidental vagaries of missing data or the sample distribution dependence of independent variable covariance.

The first row at the bottom of the figure shows the measure positions of each of the 48 gout patients and right below that the measure positions of each of the 48 gender and age matched, but not gout, patients. This provides a linear visualization of the dependent variation identified by this analysis – information seldom provided in a regression report.

On this simple linear chart, the extent to which this three-blood-chemistry measure separates these gout and "not-gout" patients is obvious. The means of the two patient groups, marked by "M's" at blood chemistry measures 45 and 60, are statistically distinct. That may be nice to publish, but clinically the visible position of each individual patient on this blood chemistry variable is far more useful.

All measures, indeed all inferences, are inevitably qualified by margins of error. We expect a region of overlap, like the one around the gout colon between 50 and 58. The vertical alignment of the "Gout?" diagnosis "(YES)"with the parenthesized Merck Manual reference values is clear evidence of the coherence between these statistical results and established reference values – an easy to see verification of validity.

Among the gout patients in Figure 2, there are two at blood chemistry measures 16 and 39. These blood chemistry measures are sufficiently low to suggest that, if these patients do have gout, it has symptoms other than blood chemistry.

Among the not gout patients there are three with blood chemistry measures in the 60's, a suspicious level according to our measure and also according to Merck.

If we use these 96 patients as current norms for this kind of gout measurement, then we can see and explain the implications of each measure position in terms of the observed odds among these 96 patients for (or against) having gout.

At blood chemistry measure 57, the observed gout odds among these patients are 6/5, just about even. At measure 64, however, observed gout odds rise to 7/1, or, if we group adjacent columns, (7+3)/(2)=5/1. These odds for the presence of gout are large enough to suggest a decision. Moving down to a measure of 52 implies gout odds of 5/8 and at measure 49 odds of only 1/3. At lower measures the observed odds against gout become overwhelming.

Even this small sample of 96 provides preliminary norms. A simple accumulation from medical records of a larger and continually growing sample will provide observed gout odds interpretations of any medical measure with increasing authority.

A final, perhaps surprising and, if so, crucial, observation clearly visible in Figure 2 and calculated in Table 2 is the non-linearity of the relationship between mg/dl chemical metrics and the metric of medical diagnosis. This non-linearity shows in the unequal medical measure distances between the integer codes which mark equal increments in chemical mg/dl.

Table 2 shows that for creatinine, the increment in diagnostic significance from code 0, marking 0.7 mg/dl, to code 1, marking 1.0 mg/dl is 22 medical units. This is .0136 mg/dl per medical measure unit. If we use 5 medical diagnosis units as our margin of error, then creatinine changes as small as .07 mg/dl could have medical implications at levels below 1 mg/dl. But the increment in medical significance from code 4, marking 1.9 mg/dl, to code 5, marking 2.2 mg/dl, is only one medical unit, or 0.3 mg/dl per medical measure unit. This means that at creatinine levels near 2 mg/dl it takes a change of 1.5 mg/dl in chemical creatinine to mean as much medically as a change of 0.07 mg/dl at levels near 1 mg/dl. The chemical mg/dl increase at codes 4 to 5 is 22 times the increase at codes 0 to 1. This implies that mg/dl increases in creatinine below 1 mg/dl are 22 times more important medically than the same size increases above 2 mg/dl. These numbers are listed in Table 2. A regression analysis is unlikely to document or even to reveal such an important finding.

Discussion

This paper shows how Rasch measurement can replace regression analysis to advantage and also provide reports far more useful to medical diagnosis. The practical implications of regression coefficients are hard to visualize, let alone understand. In addition regression coefficients are vulnerable to missing data and disturbed by sample dependent covariance. The results reported here show how the intentions of regression analysis can be better realized and more usefully reported by Rasch measurement.

This paper shows how Rasch analysis can simplify the clinician's job by constructing one simple picture from which the implications of laboratory abnormalities can be clearly seen. The illustration is based on observed relationships among laboratory findings among patients who have been diagnosed with gout by the usual methods. The analysis shows that the gout implications of corpulence can be quite distinct from blood chemistry and that cholesterol and triglycerides do not contribute useful information to a gout blood chemistry variable.

References

Acheson, R.M., & O'Brien, W.M. (1966). Dependence of serum-uric-acid on haemoglobin and other factors in the general population. *Lancet, 2*, 777-778.

Berger, L., & Yu, T-F. (1975). Renal function in gout: IV. An analysis of 524 gouty patients including long-term follow up studies. *American Journal of Medicine, 49*, 605.

Campbell, S. M. (1988). Gout: How presentation, diagnosis, and treatment differ in the elderly. *Geriatrics, 43*, 71-77.

Culleton, B. F., Larson, M. G., Kannel, W. B., & Levy, D. (1999). Serum uric acid and risk for cardiovascular disease and death: The Framingham Heart Study. *Annals of Internal Medicine, 131*, 7-13.

Evans, J. G., Prior, I. A. M., & Harvey, H. P. B. (1968). Relation of serum uric acid to body bulk, hemoglobin, and alcohol intake in two South Pacific Polynesian populations. *Annals of the Rheumatic Diseases, 27*, 319-325.

Garrick, R., Bauer, G. E., Ewan, C. E., & Meale, C. F. (1972). Serum uric acid in normal and hypertensive Australian subjects. *Australian New Zealand Journal of Medicine, 4*, 351-356.

Gibson, T., & Grahame, R. (1974). Gout and hyperlipidaemia. *Annals of the Rheumatic Diseases, 33*, 298-303.

Glynn, R. J., Campion, E. W., & Silbert, J. E. (1983). Trends in serum uric acid levels 1961-1980. *Arthritis and Rheumatism, 26*, 87-93.

Harris, M. D., Siegel, L. B., & Alloway, J. A. (1999). Gout and hyperuricemia. *American Family Physician, 59*, 925-934.

Linacre, J. M., & Wright, B. D. (2000). *WINSTEPS Rasch Analysis Computer Program*. Chicago: MESA

Murphy, M. B., Lewis, P. J., Kohner, E., Schumer, B., & Dollery, C. T. (1982). Glucose intolerance in hypertensive patients on prolonged diuretic treatment. *Lancet, 2*, 1293-1295.

Roubenoff, R. (1990). Gout and hyperuricemia. *Rheumatic Diseases Clinics of North America, 16*, 539-550.

Roubenoff, R., Klag, M. J., Mean, L. A., Liang, K. Y., Seidler, A. J., & Hochberg, M. C. (1991). Incidence and risk factors for gout in white men. *Journal of the American Medical Association, 226*, 3004-3007.

Wolfe, F. (1991). Practical therapeutics – gout and hyperuricemia. *American Family Physician, 43*, 2141-2150.

Wolfe, F., & Cathey, M. A. (1991). The misdiagnosis of gout and hyperuricemia. *Journal of Rheumatology, 18*, 1232-1234.

Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson & S. L. Hershberger (Eds.), *The New Rules of Measurement. What Every Psychologist and Educator Should Know* (pp. 65-104). Mahwah, NJ: Erlbaum.

Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis: Rasch Measurement.* Chicago: MESA Press.

Wright, B. D., & Stone, M. H. (1971). *Best Test Design: Rasch Measurement*. Chicago: MESA Press.

Wyngaarden, J. B. (1988). Gout. In J. B. Wyngaarden & L. H. Smith, Jr. (Eds.), *Cecil Textbook of Medicine* (18th ed.) (pp. 1161-1170). Philadelphia: W. B. Saunders Company.

Multiple Regression with WINSTEPS
A Rasch Solution to Regression Confusion

Benjamin D. Wright, University of Chicago

The purpose of Rasch measurement is to build and verify a useful "yardstick" – a stable, portable, reproducible instrument for making linear measures. What makes a yardstick useful is the calibration of its reference points, which mark out a visible linear metric that maintains its spacing as long as it is used in a sensible way. Just like the yardstick in your closet, a useful "yardstick" does not change the distances between its calibration marks from object to object, place to place, or time to time, as long as you apply it as sensibly as you apply the yardstick in your closet.

What follows explains how to use the Rasch measurement program, WINSTEPS (winsteps.com, Linacre, 2000) to solve multiple regression problems in a new way that avoids the sample covariance dependence and missing data problems which interfere with inferential stability.

The data used by Rasch measurement to build yardsticks can originate as any set of ordinal indicators: dichotomies, ratings, partial credits, counts, as well as any already developed metric like those in commerce and science. The way data expressed in the decimal fractions of an existing metric, like inches or mg/Dl, is entered into WINSTEPS is:

1. Recode each decimal fraction X into interval (or log interval) integer Y:

$$Y = M(X-MIN)/(MAX-MIN) + 1/2$$

For Y = 0,9 use M = 0,N <10
For Y= 00,99 use M = 00,NN <100

If an incoming metric is expected to have a ratio effect in the yardstick you are constructing, you can anticipate this by using log X instead of X in the above formula.

2. Your choice of MIN and MAX can be made locally from the smallest incoming value for MIN and the largest incoming value for MAX. Or you can choose values for MIN and MAX which are natural to their originating metric, so long as the values you choose embrace the range of incoming data.

3. Your choice of M depends on how many ordinal integer categories you wish to use for your yardstick construction. In our practice we have not encountered any situation for which M>9 was more informative than M<10, but WINSTEPS does enable you to maintain the linear articulation of your incoming decimal fractions up to 100 steps from 00 to 99 by setting M=99.

4. In order for WINSTEPS to print the original decimal fractions Y of your incoming data next to their recoded integers X, use control variable CFILE= to label each integer category X with its corresponding decimal fraction midpoint Y:

$$X = [(Y-1/2)(MAX-MIN)/M] + MIN$$
(for log interval X = exp X)

This labeling enables you to see the extent to which the scale of decimal fractions X, however linear in X, does not make a linear contribution to your new yardstick. It is often the case that the linear intervals of X do not produce linear separations among their code values Y in the new metric defined by your yardstick.

5. If any of your decimal fraction variables X have useful reference points, such as freezing at 32 degrees or normal body temperature at 98.6 degrees, you can reference your item calibration representations of these variables by pivoting the calibration of the equivalent Y integer step at that reference point.

How to Use WINSTEPS to Solve Multiple Regression Problems

1. Organize your incoming variables into three groups:

Dependent Variables = DV to be predicted by IV
Independent Variables = IV to predict DV's.
Conditional Variables = CV to condition prediction for interaction with other variables like: gender, age, culture, language, wealth . . .

2. Apply one of the following three "regression" formulations:

a. DV positioned on the IV set in terms of a *DV defined variable*.

Anchor objects (persons) at their incoming DV values. This establishes your dependent variable. Then use WINSTEPS to find the best set of IV calibrations for predicting these anchored DV values. This formulation optimizes prediction, but binds IV calibrations to DV sample dependence.

b. DV positioned on IV in terms of an *IV defined variable.*

Apply WINSTEPS to the IV set to find the best variable this IV set can define, independent of any DV. This requires a sequence of stepwise analyses by which members of the IV set are edited until a best possible IV variable has become defined (The steps are listed below).

Because the construction of the IV variable is entirely independent of DV data, this formulation enables the simultaneous evaluation of any number of DV's.

Anchor to the item/step calibrations of this best IV variable and then insert all DV's and use WINSTEPS to show how well these DV's are predicted by the IV just constructed independently of any DV distribution and also of any sample dependent covariance among the IV. This sample free construction of a single variable defined by the IV set optimizes the inferential stability of DV predictions.

c. Middle Ground *Short Cut.*

Combine all DV and IV in one WINSTEPS analysis. The result will fall between formulations (a) and (b). But they will be dominated by (b) to the extent that IV information exceeds DV information.

3. Two ways to introduce CV variables.

a. Several Separate Analyses.

For CV's with few categories, repeat Step 2 for each CV sub-group. Compare maps.

b. Sequence of Composite Analyses

Include CVs in each analysis and use person separations, fit statistics and residual analyses to expose the extent to which each CV interferes (or helps).

How to Construct a Best IV Variable

1. Item Polarity: Examine the correlations between item responses and person measures in the Item Misfit Table to identify and correct all negative relations by reversing their scoring.

2. Category Articulation: Examine the Rating Scales Structure Table to identify noisy and uninformative categories that you can improve by rescoring these categories.

3. Item Dimensionality: Examine the Item Principal Component Analysis of Response Residuals Table to find out whether there is a secondary item dimension large enough or meaningful enough to isolate.

4. Person Dimensionality: If the relative size or item content of the first item residual factor interests you, examine the Person Principal Component Analysis of Response Residuals Table to identify and evaluate the effect of this secondary dimension on person measures.

5. Variable Sharpening: Reexamine the Item Misfit Table to evaluate the effects on person separation (in the Summary Table) of deleting items with large infit mean squares (e.g.>1.3) in order to find the most efficient definition of your IV variable.

How WINSTEPS Improves on Multiple Regression MR

1. MR arithmetic and stochastic interpretation depends on normally distributed continuous linear data.

WINSTEPS accepts discrete ordinal data of any distribution and constructs linear continuous measures from them. Every analysis of raw ordinal observations requires this step to prepare for linear statistical analysis.

2. MR is vulnerable to missing data.

When rows and columns are connected, WINSTEPS conjoint additivity corrects for missing data automatically.

3. MR posits a single dimensioned DV to which the IV, whatever their own dimensions, must produce a co-linear contribution.

WINSTEPS extracts the best possible single linear dimension, which the data support and estimates continuous linear measures, standard errors and fit statistics on this dimensions for all item, step and person parameters.

4. MR regression coefficients and multiple R's are hard to interpret because they defy visualization.

WINSTEPS constructs linear measures, qualified by errors and fit statistics and reports them on linear MAPs which show, in complete detail, the positional relationship between all values of the DV in terms of all values of the IV. The resulting positional relationships are complete, easy to see and easy to interpret.

A Gout Application of WINSTEPS Regression

Table 1 shows three panels from the application of WINSTEPS Rasch regression to the medical data discussed in the article, "Rasch Measurement Instead of Regression" by Wright, Perkins and Dorsey.

The top panel lists, on the right, the eight definitions of the medical items analyzed. The next column to the left, "SCORE CORR." are response/measure correlations. For the three anchored blood measures which define the independent variable, IV, these correlations correspond to standardized regression coefficients. For the five dependent variables, DV, listed below, they correspond to the usual multiple regression prediction correlations.

Next to the left are two columns of mean square fit statistics. When these mean squares are near 1.00, they document a valid relationship among the three anchored blood chemistry IV's: Uric Acid, Urea Nitrogen and Creatinine. They also validate or invalidate the regressions on the three blood chemistry IV of the five DV's: Gout, Hypertension, Diuretics, Kidney Stones and Diabetes.

Gout does best with a prediction correlation of .61, closely followed by hypertension. Both correlations and outfit mean squares expose the failure of this three blood chemistry IV to predict kidney stones or diabetes.

The middle panel of Table 1 illustrates the IV linear coding of chemical metrics mg/dl onto 10 category integer scales, 0 to 9. It also shows the pivot marking for each blood chemistry at the Merck Manual step from "normal" to "high". The middle panel also lists the distribution of the 96 patients across the 10 levels for each IV and the average measure at each category of the new medical variable which the three blood chemistries were found to define. Since the chemical mg/dl metrics are evenly represented by the 10 categories, the non-linear distributions of these average measures is noteworthy. This non-linearity shows that the medical implications of increases in these blood chemistries are not collinear with increases in their chemical metric mg/dl. The clinical implications of a particular increment in mg/dl varies with mg/dl level. This irregularity muddies clinical evaluations of blood chemistry changes. The WINSTEPS analysis makes the specifics of this non-linearity evident and provides, instead, a new medical metric, which is linear in its clinical implications.

The bottom panel of Table 1 sums up the diagnostic implications of these analyses. The multiple regression prediction correlations, repeated from above, show that Gout at .61 is better predicted than hypertension at .51. Far more useful, however, are the measurement positions of each diagnostic indicator. The gout indicators, rounded to 48 for "No Gout" and 59 for "Gout", mark the positions on the three blood chemistry yardstick where the odds for the presence of gout shift from 1/2 at 48 to 2/1 at 59. The hypertension indicators, rounded to 49 and 59, provide a similar interpretation with respect to hypertension. At the bottom we see again, in metric form, the futility of trying to predict kidney stones or diabetes from these three blood chemistries.

When the 12 unit distance (59.44 - 47.70 = 11.74) between the gout indicators is compared to the 10 unit distance (58.83 – 48.93 = 9.90) between the hypertension indicators, we see the metric implications of their .61 > .51 multiple correlation difference. The ratio of those distances, 11.74/9.90 = 1.19, measures how much better this yardstick predicts gout than hypertension. Similar comparisons can be made among all five dependent variables.

The piece de resistance for clinical interpretation, however, is displayed in Figure 2 of "Rasch Measurement Instead of Regression" by Wright, Perkins and Dorsey. In that Figure, the position of any patient measure with respect to the "No Gout" and "Gout" indicators makes the clinical interpretation of the measure obvious. See that discussion of Figure 2 to appreciate the clinical advantage of WINSTEPS Rasch measurement "regression" analysis.

Table 1. WINSTEPS MULTIPLE REGRESSION Results

RAW SCORE	COUNT	MEASURE	ERROR	INFIT MNSQ	OUTFIT MNSQ	SCORE CORR.	ITEMS	
408	96	60.3A	.7	.83	.82	.83	URIC ACID	INDEPENDENT
325	96	62.7A	.8	.70	.75	.72	UREA NITROGEN	VARIABLES
181	96	55.3A	.8	.89	1.07	.62	CREATININE	
48	96	53.7	1.9	.80	.88	.61	GOUT	DEPENDENT
45	96	55.2	1.9	.91	1.08	.51	HyperTense	VARIABLES
22	96	66.0	2.1	1.01	.81	.39	Diuretic	Successful
6	96	79.5	3.5	1.19	3.33	-.03	KidneyStone	Unsuccessful
9	96	75.7	2.9	1.19	4.78	-.06	Diabetes	

Table 1. (continued)

SCORE VALUE	DATA COUNT	%	AVERAGE MEASURE	ITEM	DIAGNOSTIC MEASURES	
0	1	1	29.12	URIC ACID	2.1 mg/dl	
1	4	4	34.91		3.3	
2	11	11	42.12		4.5	
3	20	21	48.66		5.7	
4	21	22	55.76		6.9 normal	
5	12	13	55.28		8.1 high	55
6	14	15	61.31		9.3	
7	11	11	64.51		10.5	
8	1	1	64.67		11.7	
9	1	1	70.53		12.9	
0	1	1	29.12	UREA	0 mg/dl	INDEPENDENT
1	1	1	36.80	NITROGEN	5	
2	21	22	45.80		10	
3	41	43	52.40		15 normal	
4	18	19	57.91		20 high	58
5	6	6	62.29		25	
6	3	3	66.61		30	
7	2	2	69.44		35	
8	2	2	69.02		40	
9	1	1	73.50		45	
0	8	8	38.95	CREATININE	0.7 mg/dl	VARIABLES
1	39	41	50.63		1.0 normal	
2	28	29	55.14		1.3 high	55
3	13	14	60.61		1.6	
4	3	3	60.12		1.9	
5	1	1	67.51		2.2	
6	1	1	61.88		2.5	
7	1	1	69.75		2.8	
8	1	1	73.50		3.1	
9	1	1	71.37		3.4	
0	48	50	47.70	GOUT	DIAGNOSIS	
1	48	50	59.44	1/0= +12	R= +.61	59
0	51	53	48.93	HYPERTENSE	DIAGNOSIS	DEPENDENT
1	45	47	58.83	1/0= +10	R= +.51	59
0	74	77	51.50	DIURETIC	DIAGNOSIS	VARIABLES
1	22	23	60.52	1/0= +9	R= +.39	61
0	90	94	53.65	KidneyStone	No Diagnosis	
1	6	6	52.42	1/0= -1	R= -.03	
0	87	91	53.76	Diabetes	No Diagnosis	
1	9	9	51.77	1/0= -2	R= -.06	