TESTING STATISTICAL ASSUMPTIONS

By G. David Garson

North Carolina State University
School of Public And International Affairs

Contact:

G. David Garson, President
Statistical Publishing Associates
274 Glenn Drive
Asheboro, NC 27205 USA

Email: gdavidgarson@gmail.com
Web: www.statisticalassociates.com

Table of Contents

Overview

All statistical procedures have underlying assumptions, some more stringent than others. In some cases, violation of these assumptions will not change substantive research conclusions. In other cases, violation of assumptions will undermine meaningful research. Establishing that one's data meet the assumptions of the procedure one is using is an expected component of all quantitatively-based journal articles, theses, and dissertations.

For all volumes in the Statistical Associates "blue book" series, the assumptions of each statistical procedure are indicated in an "Assumptions" section. This volume provides a general overview of the most common data assumptions which the researcher will encounter in statistical research, as listed in the table of contents to the right.

Key Concepts and Terms

Parametric statistics

Parametric tests are significance tests which assume a certain distribution of the data (usually the normal distribution), assume an interval level of measurement, and assume homogeneity of variances when two or more samples are being compared. Most common significance tests (z tests, t-tests, and F tests) are parametric. However, it has long been established that moderate violations of parametric assumptions have little or no effect on substantive conclusions in most instances (ex., Cohen, 1969: 266-267.)

Nonparametric statistics

Nonparametric tests are ones which do not assume a particular distribution of the data. Chi-square tests are of this type.

Bootstrapped estimates

Bootstrapped estimates are a nonparametric approach which bases standard errors for any statistic not on assumptions about, say, the normal curve, but on the empirical distribution arising from repeated sampling from the researcher's own dataset. In doing so, however, bootstrapping changes the meaning of the p significance value. With random sampling from a normal or known distribution, p is the probability of getting a result as strong or stronger than the observed result just by chance of taking another random sample from the population. With bootstrapping, p is the probability of getting a result as strong or stronger than the observed result just by chance of taking another sample of n-1 from the researcher's sample.

Assumptions

SOUND MEASUREMENT

Descriptive statistics

All forms of statistical analysis assume sound measurement, relatively free of coding errors. It is good practice to run descriptive statistics on one's data so that one is confident that data are generally as expected in terms of means and standard deviations, and there are no out-of-bounds entries beyond the expected range.

Avoiding attenuation

When the range of the data is reduced artificially, as by classifying or dichotomizing a continuous variable, correlation is attenuated, often leading to underestimation of the effect size of that variable.

Attenuation is illustrated in the figure on the following page.

ATTENUATION

In the partial data table below, age, education, and income were duplicated as binned variables with only three categories each.

	age	educ	Income	agecat	educat	incomecat
1	60	12	120000.00	3	1	2
2	27	17	100000.00	1	3	1
3	36	12	120000.00	2	1	2
4	21	13	100000.00	1	2	1
5	35	16	120000.00	1	3	2
6	33	16	110000.00	1	3	1
7	43	12	120000.00	2	1	2
8	29	13	120000.00	1	2	2

In the correlation output below, we see the correlation is reduced for the binned variables compared to the original continuous variables. For the age*income correlation, the reduction is enough to shift the correlation from significance to nonsignificance.

Correlations

		Age of respondent	Highest year of school completed	Income	Age of respondent (Binned)	Highest year of school completed (Binned)	Income (Binned)
Age of respondent	Pearson Correlation	1	-.175**	-.044*	.907**	-.127**	-.076**
	Sig. (2-tailed)		.000	.028	.000	.000	.000
	N	2828	2816	2501	2828	2816	2501
Highest year of school completed	Pearson Correlation	-.175**	1	.310**	-.126**	.846**	.324**
	Sig. (2-tailed)	.000		.000	.000	.000	.000
	N	2816	2820	2495	2816	2820	2495
Income	Pearson Correlation	-.044*	.310**	1	.000	.219**	.706**
	Sig. (2-tailed)	.028	.000		.998	.000	.000
	N	2501	2495	2503	2501	2495	2503
Age of respondent (Binned)	Pearson Correlation	.907**	-.126**	.000	1	-.096**	-.018
	Sig. (2-tailed)	.000	.000	.998		.000	.373
	N	2828	2816	2501	2828	2816	2501
Highest year of school completed (Binned)	Pearson Correlation	-.127**	.846**	.219**	-.096**	1	.249**
	Sig. (2-tailed)	.000	.000	.000	.000		.000
	N	2816	2820	2495	2816	2820	2495
Income (Binned)	Pearson Correlation	-.076**	.324**	.706**	-.018	.249**	1
	Sig. (2-tailed)	.000	.000	.000	.373	.000	
	N	2501	2495	2503	2501	2495	2503

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Note that binning may decrease correlations, increase them, or leave them the same. However, the most common (expected) effect is to attenuate (decrease) correlation.

Avoiding tautological correlation

When the indicators for latent variable A conceptually overlap with or even include one or more of the indicators for latent variable B, definitional overlap confounds the correlation of A and B. This is particularly problemetic when indicators on the independent side of the equation conceptually overlap with indicators on the dependent side of the equation. Avoiding tautological correlation is the issue of establishing discriminant validity, discussed in the separate "blue book" volume on validity.

PROPER MODEL SPECIFICATION

Specification of a model

Specification refers to not omitting significant causal variables or including correlated but causally extraneous ones, and also to correctly indicating the direction of arrows connecting the variables in the model. When a misspecification error is corrected by changing the model, all parameter estimates in the model are subject to change, not only in magnitude, but sometimes even in direction. There is no statistical test for misspecification. A good literature review is important in identifying variables which need to be specified.

As a rule of thumb, the lower the overall effect (ex., R^2 in multiple regression, goodness of fit in logistic regression), the more likely it is that important variables have been omitted from the model and that existing interpretations of the model will change when the model is correctly specified. The specification problem is lessened when the research task is simply to compare models to see which has a better fit to the data, as opposed to the purpose being to justify one model and then assess the relative importance of the independent variables.

CELL SIZE ADEQUACY

Adequate cell size

Adequate cell count is an assumption of any procedure which uses Pearson chi-square or model likelihood chi-square (deviance chi-square) in significance testing when categorical predictors are present. This includes but is not limited to chi-

square testing of crosstabulation, loglinear analysis, binomial logistic regression, multinomial logistic regression, ordinal regression, and general or generalized linear models of the same.

Factor space

Factor space is the set of cells which are generated by a crosstabulation of the categorical dependent with all the categorical factors but not the continuous covariates. In SPSS, select Analyze, Descriptives, Crosstabs; enter the categorical dependent as the column variable and the first categorical predicator as the row variable; enter additional facts as a sequence of "Levels". The SPSS syntax for a categorical dependent with five predictor factors will be of the form:

```
CROSSTABS
  /TABLES=dependentvar BY factor1 BY factor2 BY factor3 BY
factor4 BY factor5
```

Procedures such as logistic regression will report the number of 0 cells, but if covariates are present, this number is not the number of 0 cells in factor space and should be ignored. One could re-run the model without covariates to get the correct number, but SPSS will still not report the number of cells not over 5 (see below).

Cell count rule of thumb

The widely accepted rule of thumb is that no cell in factor space should be 0 and 80% of cells should be greater than 5. Any cell with zero or low count represents a type of subject about which there are too few subjects to generalize. For example, in a study of political parties, crosstabulation might reveal that there were only 3 African-American black women under 21 who voted Republican. While the researcher might wish to generalize about age or race or gender in relation to vote, any such generalization should note African-American black women under 21 who voted Republican as an exception. If there are only one or two such cells, the exception may be noted by the researcher in a footnote. If there are many such cells in factor space, there will be many exceptions to generalization. When there are any 0 cells or more than 20% of cells under 6, the researcher should not generalize at all. Note that this logic applies to all situations involving categorical variables.

Cell size and sample size

Required sample size for given procedures is discussed in other modules, but note that even large sample size does not guarantee adequate cell size.

Adequate cell size for other procedures

Adequate cell size is a problem for any categorical analysis, not just those involving chi-square. However, statistics has traditions and fads, like all human endeavors. It is traditional for texts to cite the adequate cell size issue for some procedures and not for others. For instance, few discussions of multiple regression cite the adequate cell size problem, based on a tradition going back to when multiple regression was used only with continuous variables.

Almost any statistic (R-squared, for ex., in OLS regression) presents itself as uniformly true for everyone in the sample. However, if race were a variable and all the cells for African-Americans had 0 or most of them were <5, all researchers would agree that not much could be said about that racial group. In general, factor space is the k-way table the researcher gets when she or he asks for a crosstabulation of all k factors (categorical variables) in the analysis).

Technically, what the researcher should state is that R-squared is, say, .50 for the sample, except that cannot be stated that for zero and low-count cells in the factor space (e.g., for high-income Protestant black women, low-income Jewish Asian men, and other combinations where there are too few observations). Presenting a perhaps long list of exceptions is not done in practice. Instead the researcher is considered justified to make the across-the-board statement that "R-squared is .50 for the sample" if there are few enough zero and low-count cells in the factor space. "Few enough" is widely accepted to be no zero-count cells in factor space and 80% of the cells > 5 count. This reasoning applies to almost all statistical procedures using categorical variables because the researcher cannot generalize to types of individuals for which he or she has no or few examples.

DATA LEVEL

Data level requirements

Measurement level requirements vary by statistical procedure but many statistical procedures require an interval or ratio level of measurement. In past

decades it was common for social scientists to use ordinal data in parametric procedures requiring normally distributed continuous data. Today, with ordinal regression and many other ordinal-compatible techniques, this is no longer consensually acceptable for confirmatory analysis. However, combining multiple ordinal items into an ordinal scale, whose reliability is validated, is generally accepted for use with parametric procedures.

Nonetheless, it is still not uncommon in social science to utilize dichotomies and ordinal data, such as Likert scale data, even in procedures which technically require interval-level data. Dichotomies are often included in statistical models (ex., regression models) provided the split is less than 90:10. Ordinal variables are often included in statistical models provided the normality of error terms may be demonstrated, as discussed below. Some researchers require the ordinal scale to have five or more values.

Violations of data level assumptions mean that actual standard error will be greater than the computed standard error and significance will be overestimated (that is, the chance of Type I error is greater than computed). Using ordinal data where inappropriate is a form of measurement error, which tends to attenuate (lessen) effect sizes. Therefore if the researcher has a positive finding, it is probably understated. If there is a negative finding, however, the researcher has lost power and may arrive at an erroneous conclusion.

Test of near-metricity of ordinal data. Multidimensional scaling provides one method of testing for near-metricity of ordinal data, as discussed in the separate Statistical Associates "Blue Book" volume on “Multidimensional Scaling”. Data levels are also discussed further in a separate Statistical Associates "Blue Book" volume on that “Data Levels”.

CENTERED DATA

Centering data

Centering is not an assumption for any given statistical technique but it is often strongly recommended and without it, coefficients may lack real-world meaning. Centering is subtracting the mean from predictor variables. Centering is for continuous variables, having the effect that "controlling for other variables in the model" means holding them at the means rather than at 0. The former, of course,

is meaningful while the latter is often out-of-range. Also, centering may reduce multicollinearity.

Grand mean centering vs. group mean centering

Centering as discussed above is "grand mean centering." Group mean centering involves subtracting the group mean from cases that are in that group. Unlike grand mean centering, group mean centering changes the meaning of the variables. For example, grand mean centered income is just income rescaled so 0 is mean income. For states as groups, group mean centered income is income deviations from state average incomes. Group mean centering is uncommon but appropriate if the new meaning is indeed the focus of research.

Categorical data and centering

For categorical variables, by default the reference category is the highest-coded category. "Controlling for other variables in the model" means holding them at their reference category. Categorical variables usually are not centered. It is recommended that categorical variables be coded so that the category of greatest interest is the reference category. Using a residual category or a category with few cases undermines interpretation of the data.

UNIDIMENSIONALITY

Testing unidimensionality

If the researcher is attempting to measure a construct with multiple indicator variables, then the researcher must demonstrate that the items measure the same thing because lack of unidimensionality is a form of measurement error. Measurement error attenuates correlation and increases standard error. There are several methods of testing unidimensionality, with varying meanings of and stringency for testing for unidimensionality. Some of these methods are:

- *Cronbach's alpha*. Perhaps the most commonly used test, Cronbach's alpha is a measure of the intercorrelation of items. If alpha is greater than or equal to .8, then the items are considered unidimensional for confirmatory purposes and may be combined in an index or scale. Some researchers use the less stringent cutoff of .7, while others (including this author) consider the .7<=alpha<.8 range to be suitable for exploratory purposes only. The

most lenient authors consider the .6<=alpha<.7 range suitable for exploratory purposes while others disparage this practice. Cronbach's alpha is found in SPSS, among other places, under Analyze, Scale, Reliability Analysis.

- *Factor analysis*. Principal components factor analysis is performed on all the indicators for all the constructs in the study. Indicators should have higher factor loadings on their own constructs than on other constructs. Some researchers also require that loadings be higher than some absolute cutoff value, such as .3. Some researchers also require that indicators not crossload on factors not their own (ex., that all loadings other than their own factor be below some absolute cutoff value, such as .3). Factor analysis is found in SPSS under Analyze, Data Reduction, Factor.

- *Confirmatory factor analysis in structural equation modeling*. The first step of structural equation modeling is confirmatory factor analysis, where the measurement model is assessed separately from the structural model. If goodness of fit measures for the measurement model are acceptable, the researcher concludes that the indicators adequately measure the intended constructs. It is also possible to assess the unidimensionality of a concept by comparing models in structural equation modeling Confirmatory factor analysis can be implemented through AMOS, distributed by SPSS as its structural equation modeling package.

- *Guttman scaling and other forms of scaling*. In Guttman, proximity, Mokken, and certain other types of scales, indicator variables are tested to see if they form a certain relationship to each other, such that the researcher is justified in putting the items in the same scale. The relationships, and hence the meanings of unidimensionality, differ by scale type. Note that the most common form of scale, the Likert scale, does not normally involve testing for unidimensionality. This is discussed in more detail in the section on scales and standard measures. Guttman scaling is found in SPSS under Analyze, Scale, Reliability Analysis.

Note that a set of items may be considered to be unidimensional using one of the methods above, even when another method would fail to find statistical justification in considering the items to measure a single construct. For instance, it

would be quite common to find that items in a large Guttman scale would fail to load on a single factor using factor analysis as a method. Finding satisfactory model fit in SEM would not assure that the Cronbach's alpha criterion was met. The researcher must decide on theoretical grounds what definition and criterion for unidimensionality best serves his or her research purpose. However, *some* method must always be used before proceeding to use multiple indicators to measure a concept.

NORMALITY

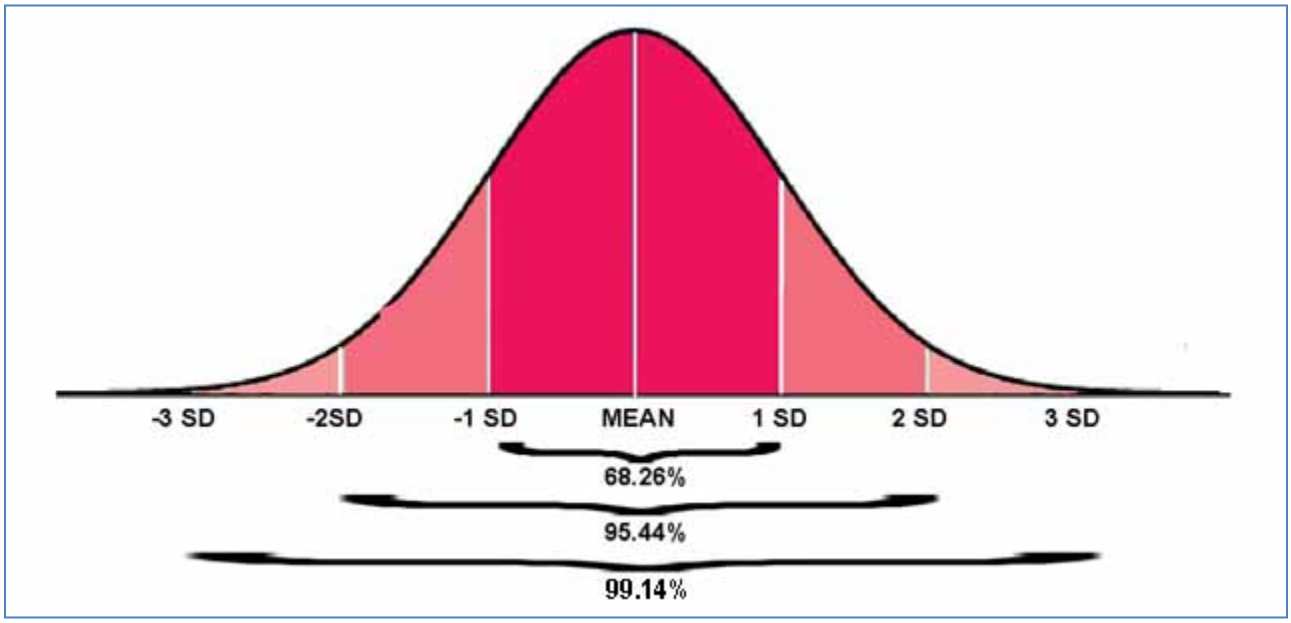
Normal distribution

A normal distribution is assumed by many statistical procedures. Various transformations are used to correct non-normally distributed data. Correlation, least-squares regression, factor analysis, and related linear techniques are relatively robust against non-extreme deviations from normality provided errors are not severely asymmetric (Vasu, 1979). Severe asymmetry might arise due to strong outliers. Log-linear analysis, logistic regression, and related techniques using maximum likelihood estimation are even more robust against moderate departures from normality (cf. Steenkamp & van Trijp, 1991: 285). Likewise, Monte Carlo simulations show the t-test is robust against moderate violations of normality (Boneau, 1960).

Normal distributions take the form of a symmetric bell-shaped curve. The *standard* normal distribution is one with a mean of 0 and a standard deviation of 1. *Standard scores*, also called z-scores or standardized data, are scores which have had the mean subtracted and which have been divided by the standard deviation to yield scores which have a mean of 0 and a standard deviation of 1. Normality can be visually assessed by looking at a histogram of frequencies, or by looking at a normal probability plot output by most computer programs.

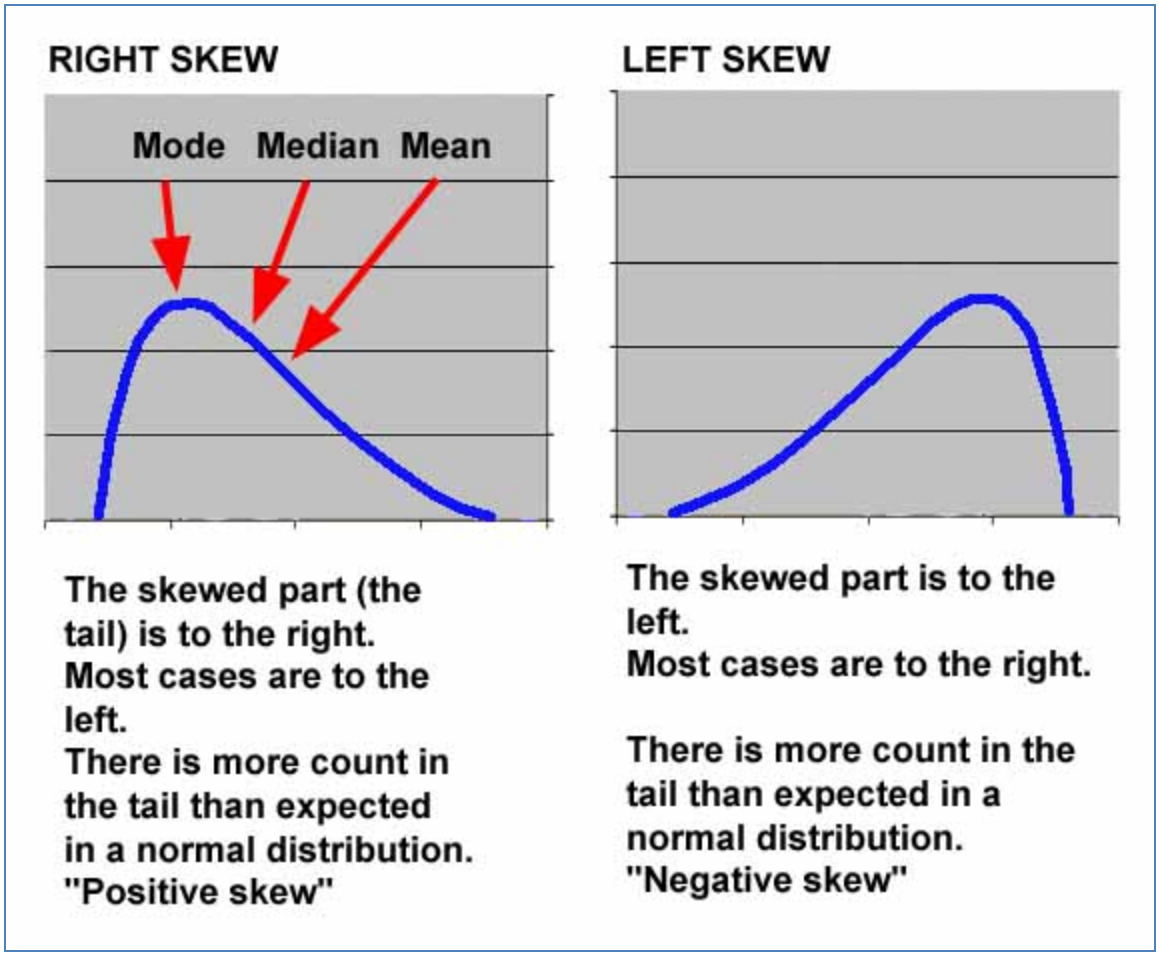
The area under the normal curve represents probability: 68.26% of cases will lie within 1 standard deviation of the mean, 95.44% within 2 standard deviations, and 99.14% within 3 standard deviations. Often this is simplified by rounding to say that 1 s.d. corresponds to 2/3 of the cases, 2 s.d. to 95%, and 3 s.d. to 99%. Another way to put this is to say there is less than a .05 chance that a sampled case will lie outside 2 standard deviations of the mean, and less than .01 chance

that it will lie outside 3 standard deviations. This statement is analogous to statements pertaining to significance levels of .05 and .01, for two-tailed tests. .



Skew

Skew is the tilt (or lack of it) in a distribution. The more common type is right skew, where the tail points to the right. Less common is left skew, where the tail is points left. A common rule-of-thumb test for normality is to run descriptive statistics to get skewness and kurtosis, then divide these by the standard errors. Skew should be within the +2 to -2 range when the data are normally distributed. Some authors use +1 to -1 as a more stringent criterion when normality is critical. In SPSS, one of the places skew is reported is under Analyze, Descriptive Statistics, Descriptives; click Options; select skew.

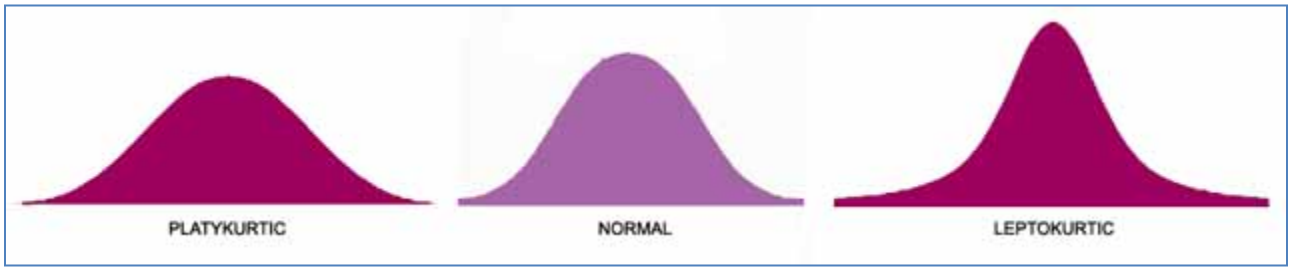


Negative skew is right-leaning, positive skew is left-leaning. For each type of skew, the mean, median, and mode diverge, so all three measures of central tendency should be reported for skewed data. Box-Cox transformation may normalize skew. Right-skewed distribution may fit power, lognormal, gamma, Weibull, or chi-square distributions. Left-skewed distributions may be recoded to be right-skewed. (Note: there is confusion in the literature about what is "right" or "left" skew, but the foregoing is the most widely accepted labeling.)

Kurtosis

Kurtosis is the peakedness of a distribution. A common rule-of-thumb test for normality is to run descriptive statistics to get skewness and kurtosis, then use the criterion that kurtosis should be within the +2 to -2 range when the data are normally distributed (a few authors use the more lenient +3 to -3, while other authors use +1 to -1 as a more stringent criterion when normality is critical). Negative kurtosis indicates too many cases in the tails of the distribution. Positive

kurtosis indicates too few cases in the tails. Note that the origin in computing kurtosis for a normal distribution is 3 and a few statistical packages center on 3, but the foregoing discussion assumes that 3 has been subtracted to center on 0, as is done in SPSS and LISREL. The version with the normal distribution centered at 0 is *Fisher kurtosis*, while the version centered at 3 is *Pearson kurtosis*. SPSS uses Fisher kurtosis. *Leptokurtosis* is a peaked distribution with "fat tails", indicated by kurtosis > 0 (for Fisher kurtosis, or > 3 for Pearson kurtosis). *Platykurtosis* is less peaked "thin tails" distribution, with a kurtosis value < 0 (for Fisher kurtosis, or < 3 for Pearson kurtosis).



Various transformations are used to correct kurtosis: cube roots and sine transforms may correct negative kurtosis. In SPSS, one of the places kurtosis is reported is under Analyze, Descriptive Statistics, Descriptives; click Options; select kurtosis.

Dichotomies

By definition, a dichotomy is not normally distributed. Many researchers will use dichotomies for procedures requiring a normal distribution as long as the split is less than 90:10. Dichotomies should not be used as dependents in procedures, such as OLS regression, which assume a normally distributed dependent variable.

Shapiro-Wilk's W test

This test is a formal test of normality offered in the SPSS EXAMINE module or the SAS UNIVARIATE procedure. This is the standard test for normality. In SPSS, select Analyze, Descriptive statistics, Explore. Note the Explore menu choice pastes the EXAMINE code. Click the Plots button and check "Normality plots with tests." Output like that below is generated:

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
StdEduc	.017	932	.200*	.999	932	.960

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

For a given variable, W should not be significant if the variable's distribution is not significantly different from normal, as is the case for StdEduc in the illustration above. W may be thought of as the correlation between given data and their corresponding normal scores, with W = 1 when the given data are perfectly normal in distribution. When W is significantly smaller than 1, the assumption of normality is not met. Shapiro-Wilk's W is recommended for small and medium samples up to n = 2000. For larger samples, the Kolmogorov-Smirnov test is recommended by SAS and others.

Kolmogorov-Smirnov D test or K-S Lilliefors test

This test is an alternative test of normality for large samples, available in SPSS EXAMINE and SAS UNIVARIATE. This test is also found in SPSS under Analyze, Descriptive Statistics, Explore, Plots when one checks "Normality plots with tests." Output is illuatrated above. Kolmogorov-Smirnov D is sometimes called the *Lilliefors test* as a correction to K-S developed by Lilliefors is now normally applied. SPSS, for instance, automatically applies the Lilliefors correction to the K-S test for normality in the EXAMINE module (but not in the NONPAR module). This test, with the Lilliefors correction, is preferred to the chi-square goodness-of-fit test when data are interval or near-interval. When applied without the Lilliefors correction, K-S is very conservative: that is, there is an elevated likelihood of a finding of non-normality. Note the K-S test can test goodness-of-fit against any theoretical distribution, not just the normal distribution. Be aware that when sample size is large, even unimportant deviations from normality may be technically significant by this and other tests. For this reason it is recommended to use other bases of judgment, such as frequency distributions and stem-and-leaf plots.

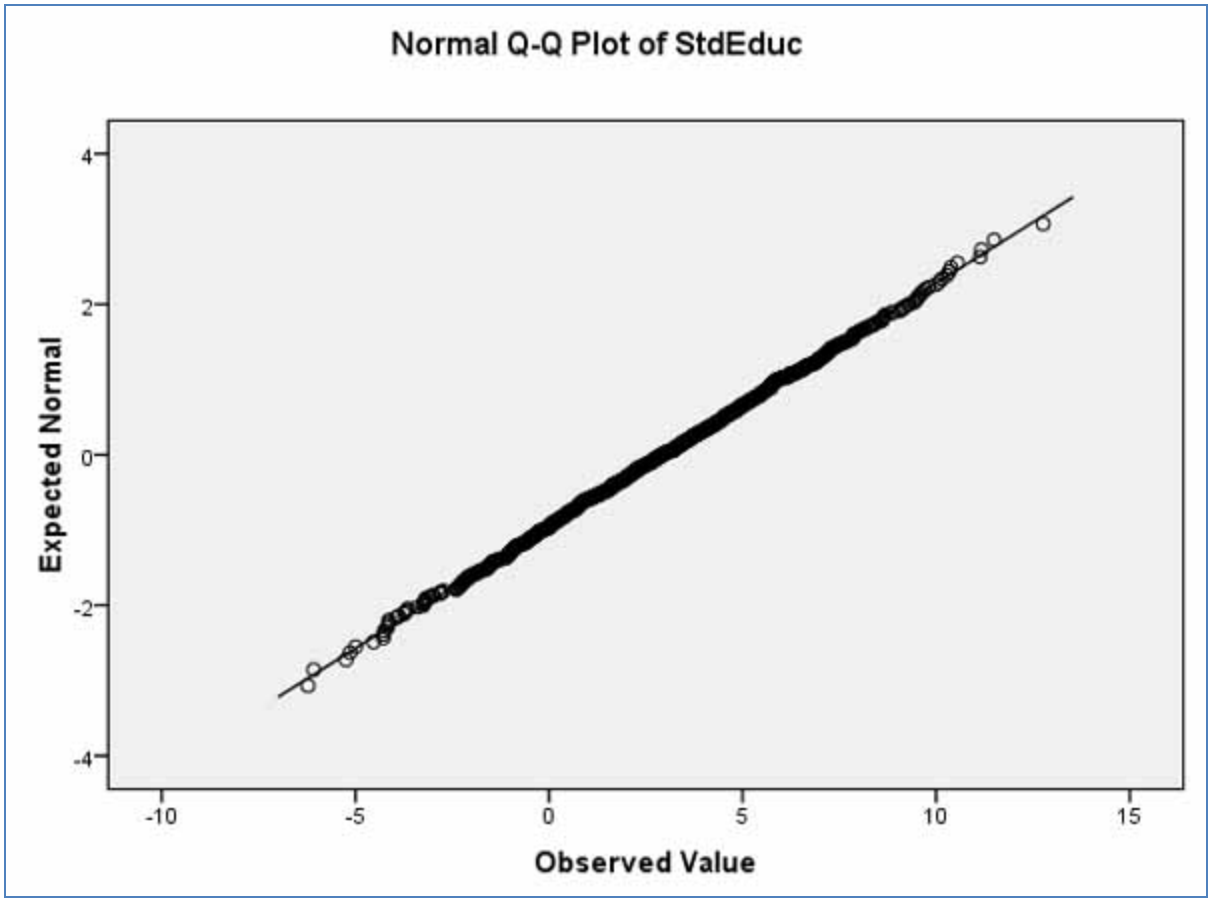
Graphical methods of assessing normality

A histogram of a variable shows rough normality, and a histogram of residuals, if normally distributed, is often taken as evidence of normality of all the variables.

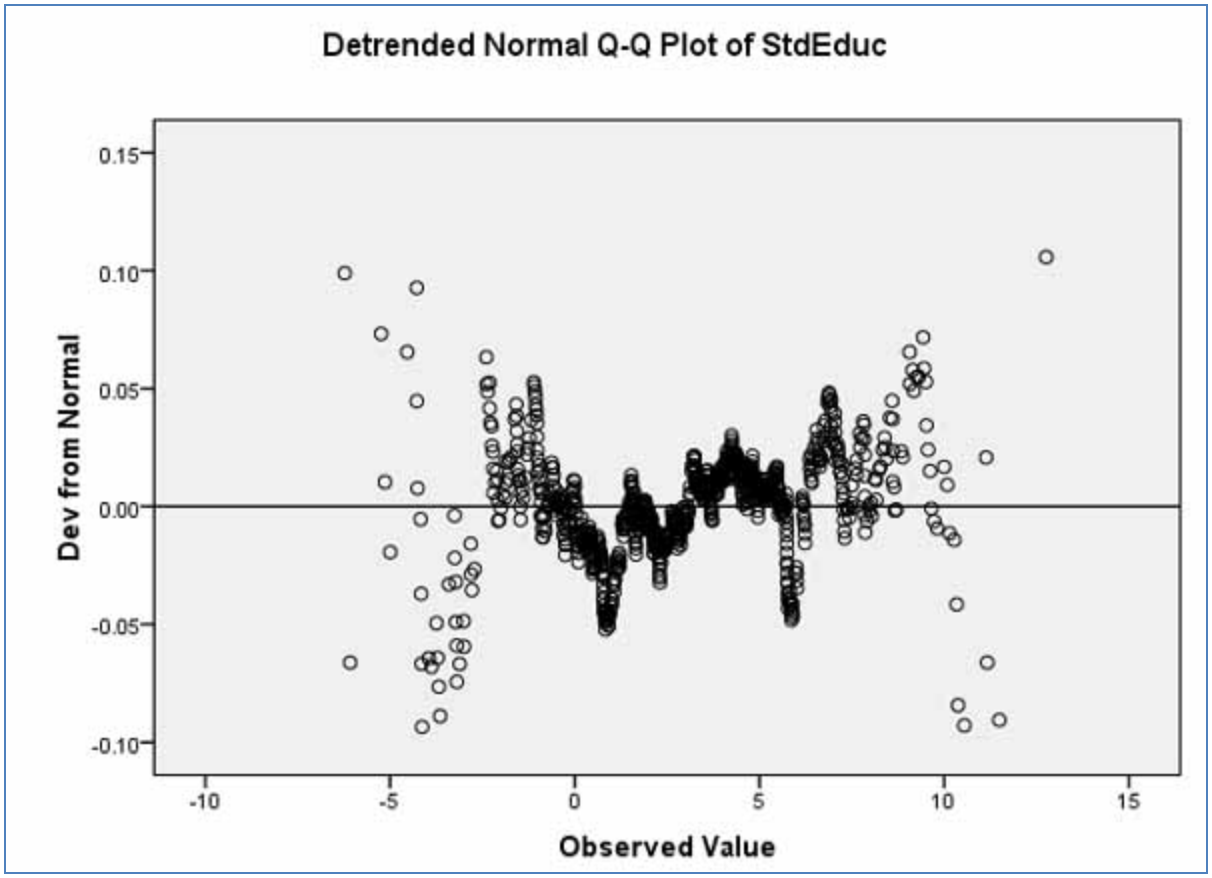
A graph of empirical by theoretical cumulative distribution functions (cdf's) simply shows the empirical distibution as, say, a dotted line, and the hypothetical distribution, say the normal curve, as a solid line.

A P-P plot is found in SPSS under Graphs, P-P plots. One may test if the distribution of a given variable is normal (or beta, chi-square, exponential, gamma, half-normal, Laplace, Logistic, Lognormal, Pareto, Student's t, Weibull, or uniform). he P=P plot plots a variable's cumulative *proportions* against the cumulative proportions of the test distribution.The straighter the line formed by the P-P plot, the more the variable's distribution conforms to the selected test distribution (ex., normal). Options within this SPSS procedure allow data transforms first (natural log, standardization of values, difference, and seasonally difference).

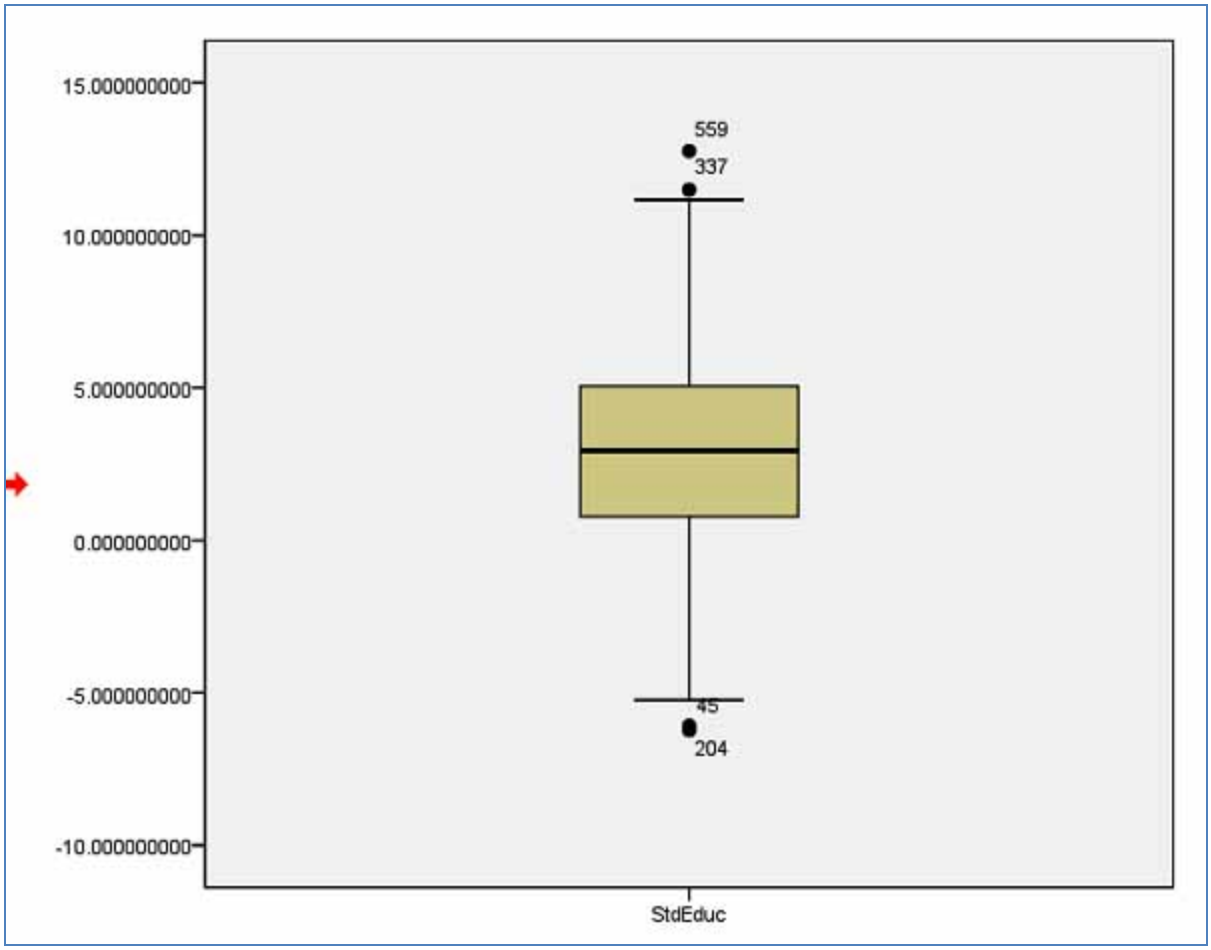
A quantile-by-quantile or Q-Q plot forms a 45-degree line when the observed values are in conformity with the hypothetical distribution. Q-Q plots plot the *quantiles* of a variable's distribution against the quantiles of the test distribution.From the SPSS menu, select Graphs, Q-Q. The SPSS dialog box supports testing the following distributions: beta, chi-square, exponential, gamma, half-normal, Laplace, Logistic, Lognormal, normal, pareto, Student's t, Weibull, and uniform. Q-Q plots are also produced in SPSS under Analyze, Descriptive Statistics, Explore, Plots when one checks "Normality plots with tests."



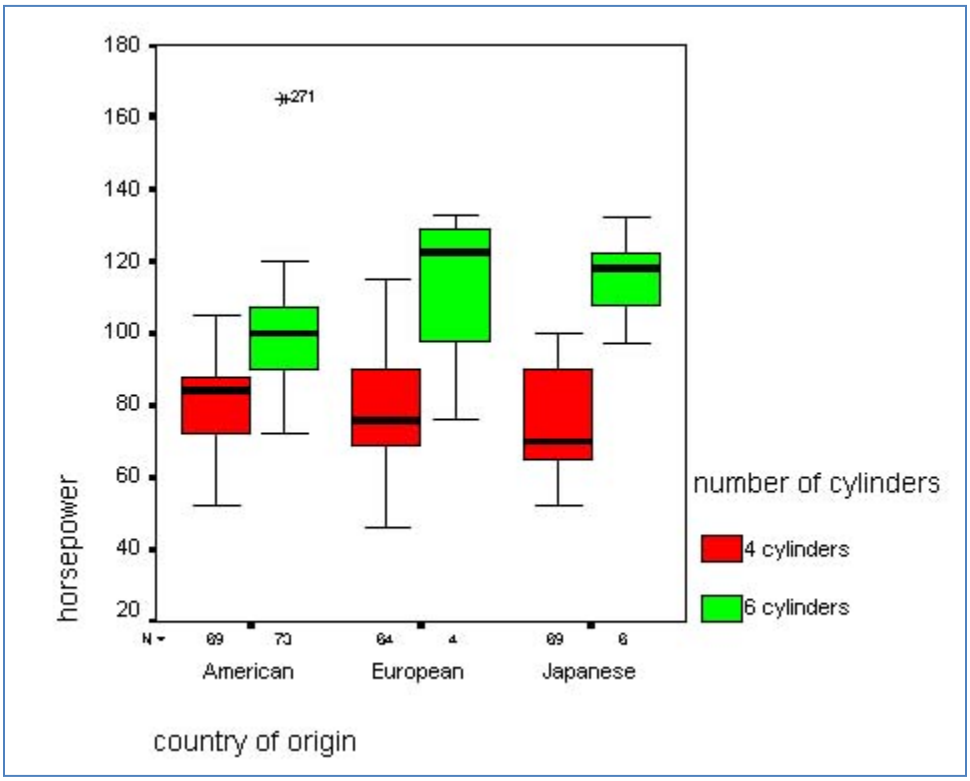
A detrended Q-Q plot, obtained in the same way in SPSS, provides similar information. If a variable is normally distributed, cases in the detrended Q-Q plot should cluster around the horizontal 0 line representing 0 standard deviations from the 45=degree line seen in the non-detrended Q-Q plot above. The detrended Q-Q plot is useful for spotting outliers. For the illustration below, however, there are no outliers in the sense that there are no cases more than +/- .12 standard deviations away. Cases more than +/- 1.96 standard deviations away are outliers at the .95 confidence level.



Boxplot tests of the normality assumption. Outliers and skewness indicate non-normality, and both can be checked with boxplots. The SPSS boxplot output option (also under Analyze > Descriptive Statistics > Explore, “Plots” button, check "Normality plots with tests"). For a single variable being tested for normality, a box plot is a chart with that variable on the X axis and with the Y axis representing its spread of values (in the illustration below, the values of standardized education, StdEduc). Inside the graph, for the given variable, the height of the rectangle indicates the spread of the values for the variable. The horizontal dark line within the rectangle indicates the mean. In the illustration it is 0 since zero is the mean for standardized variables, which StdEduc is. If most of the rectangle is on one side or the other of the mean line, this indicates the dependent is skewed (not normal). Further out than the rectangle are the "whiskers," which mark the smallest and largest observations which are not outliers (defined as observations greater than 1.5 inter-quartile ranges [IQR's = boxlengths] from the 1st and 3rd quartiles). Outliers are shown as numbered cases beyond the whiskers. (Note you can display boxplots for two factors (two independents) together by selecting Clustered Boxplots from the Boxplot item on the SPSS Graphs menu.)



- To take a more complex example, a box plot can also be a chart in which categories of a categorical independent or of multiple independents are arrayed on the X axis and values of an interval dependent are arrayed on the X axis. In the example below there are two categorical independents (country of car manufacture, number of cylinders) predicting the continuous dependent variable horsepower.
- Inside the graph, for each X category, will be a rectangle indicating the spread of the dependent's values for that category. If these rectangles are roughly at the same Y elevation for all categories, this indicates little difference among groups. Within each rectangle is a horizontal dark line, indicating the mean. If most of the rectangle is on one side or the other of the mean line, this indicates the dependent is skewed (not normal) for that group (category). Whiskers and outliers are as described above for the one-variable case.



Resampling

Resampling, which is a form of bootstrapping, is a way of doing significance testing while avoiding parametric assumptions like multivariate normality. The assumption of multivariate normality is violated when dichtomous, dummy, and other discrete variables are used. In such situations, where significance testing is appropriate, researchers may use a resampling method.

Normalizing Transformations

Various transformations are used to correct skew:

- Square roots, logarithmic, and inverse (1/x) transforms "pull in" outliers and normalize right (positive) skew. Inverse (reciprocal) transforms are stronger than logarithmic, which are stronger than roots.

- To correct left (negative) skew, first subtract all values from the highest value plus 1, then apply square root, inverse, or logarithmic transforms.

- For power and root transforms, finer adjustments can be obtained by adding a constant, C, where C is some small positive value such as .5, in the transform of X: $X' = (X + C)^P$. When this researcher's data contain zero values, the transform using C is strongly recommended over straight transforms (ex., SQRT(X+.5), not SQRT(X)), but the use of C is standard practice in any event. Values of P less than one (roots) correct right skew, which is the common situation (using a power of 2/3 is common when attempting to normalize). Values of P greater than 1 (powers) correct left skew. For right skew, decreasing P decreases right skew. Too great reduction of P will overcorrect and cause left skew. When the best P is found, further refinements can be made by adjusting C. For right skew, for instance, subtracting C will decrease skew.

- Logs vs. roots: logarithmic transformations are appropriate to achieve symmetry in the central distribution when symmetry of the tails is not important; square root transformations are used when symmetry in the tails is important; when both are important, a fourth root transform may work (fourth roots are used to correct extreme skew).

- Logit and probit transforms. Schumacker & Lomax (2004: 33) recommend probit transforms as a means of dealing with skewness. See Lipsey & Wilson (2001: 56) for discussion of logit and probit transforms as a means of transforming dichotomous data as part of estimating effect sizes.

- Percentages may be normalized by an arcsine transformation, which is recommended when percentages are outside the range 30% - 70%. The more observations outside this range or the closer to 0% and/or 100%, the more normality is violated and the stronger the recommendation to use arcsine transformation. However, arcsine transformation is not effective when a substantial number of observations are 0% or 100%, or when sample size is small. The usual arcsine transformation is $p' = \arcsin(\sqrt{p})$, where p is the percentage or proportion.

- Poisson distributions may be normalized by a square root transformation.

- Other strategies to correct for skew include collapsing categories and dropping outliers.

Warnings. Transformations should make theoretical sense. Often, normalizing a dichotomy such as gender will not make theoretical sense. Also note that as the log of zero is undefined and leads to error messages, researchers often add some arbitrary small value such as .001 to all values in order to remove zeros from the dataset. However, the choice of the constant can affect the significance levels of the computed coefficients for the logged variables. If this strategy is pursued, the researcher should employ sensitivity analysis with different constants to note effects which might change conclusions.

Transforms in SPSS: Select Transform - Compute - Target Variable (input a new variable name) - Numeric Expression (input transform formula)

Box-Cox Transformations of Dependent Variables

Box & Cox proposed a maximum likelihood method in 1964 for determining the optimal power transform for purposes of normalization of data. Power transformations of dependent variables were advanced to remedy model lack of normal distribution, lack of homogeneity of variances, and lack of additivity. In a regression context, Box-Cox transformation addresses the problem of non-normality, indicated by skewed residuals (the transformation pulls in the skew) and/or by lack of homoscedasticity of points about the regression line. In an ANOVA context, the Box-Cox transformation addresses the problem of lack of homogeneity of variances associated with the correlation of variances with means in the groups formed by the independent factors and indicated by skewed distributions within the groups (the transformation reduces the correlation).

Procedure. In general, the Box-Cox procedure is to (1) Divide the independent variable into 10 or so regions; (2). Calculate the mean and s.d. for each region; (3). Plot log(s.d.) vs. log(mean) for the set of regions; (4). If the plot is a straight line, note its slope, b, then transform the variable by raising the dependent variable to the power (1 - b), and if b = 1, then take the log of the dependent variable; and (5) if there are multiple independents, repeat steps 1 - 4 for each independent variable and pick a b which is the range of b's.

Lambda. In practice, computer packages apply an iterative maximum-likelihood algorithm to compute lambda, a Box-Cox parameter used to determine the exact power transformation which will best de-correlate the variances and means of the groups formed by the independent variables. As a rule of thumb, if lambda is 1.0, no transformation is needed. A lambda of +.5 corresponds to a square root transform of the dependent variable; lambda of 0 corresponds to a natural log transform; -.5 corresponds to a reciprocal square root transform; and a lambda of -1.0 corresponds to a reciprocal transform. The Box-Cox transformation is not yet supported in SPSS.

Computer packages. However, applying the Box-Cox transformation in SPSS and SAS is discussed in detail in Jason W. Osborne's article, "Improving your data transformations: Applying the Box-Cox transformation", *Practical Assessment, Research & Evaluation* 15(12), October, 2010.

References. See Box, G. E. P. and D. R. Cox (1964). An analysis of transformations. *Journal of the Royal Statistical Society*, B, 26, 211-234; see also Maddala, G. S. (1977). *Econometrics*. New York: McGraw-Hill. (page 315-317); or Mason, R., L., R. F. Gunst, and J. L. Hess (1989). *Statistical design and analysis of experiments with applications to engineering and science.* New York: Wiley.

MULTIVARIATE NORMALITY

Multivariate normality

Multivariate normality is when each variable under consideration is normally distributed with respect to each other variable. Multiple analysis of variance (MANOVA), for instance, requires multivariate normality in the response (dependent) variables. Structural equation modeling and certain other procedures also assume multivariate normality.

Mardia's statistic

Mardia's statistic is a test for multivariate normality. Based on functions of skewness and kurtosis, Mardia's PK should be less than 3 to assume the assumption of multivariate normality is met. PRELIS (companion software for LISREL) outputs PK. SPSS does not yet support Mardia's PK.

Univariate screening for multivariate normality

As a "quick and dirty" method, some researchers test for normality of each variable in the model, then assume if all are normally distributed, multivariate normality exists. This approach does not assure correct conclusions.

Bivariate screening for multivariate normality

A bivariate scatterplot for any pair of variables in the model should yield and oval-shaped array of points if both variables are linearly related and normally distributed. While a step up, this is considered an exploratory approach.

Residuals test.

One approach is to regress each variable in the model on all other variables in the model, then save the residuals. If all the residual variables are normally distributed (ex., by Q-Q plots, acceptable skew and kurtosis, etc.), then it is assumed that the data are multivariate normal.

OUTLIERS

Outlying observations can radically alter the outcome of analysis and are also violations of normality, so dropping outliers may be appropriate. On the other hand, dropping outliers can bias the researcher's results. As a general principle, dropping outliers is justified only if the data are bad for reasons given below, or if outlying cases will be analyzed separately because a different model dynamic applies, or if they represent a trivial proportion of the data.

Outliers arise from varied causes, requiring different courses of action:

- *Errors of data entry*: proofread your data for out-of-range entries and erroneous, discrepant, or dishonest entries. Some instruments contain crosschecks for this purpose, such as asking age in one item and birth year in another, so discrepancies might be identified.

- *Not defining missing values*: check in SPSS or other statpacks to make sure 'don't know', 'not home', and other missing values are not being treated as real values. .

- *Unintended sampling*: eliminate non-population members from the sample (ex., eliminate unintentionally sampled out-of-town house guests from a sample for the population of city residents).

- *Separate models.* The researcher may choose to analyze extreme cases separately. It is valid to remove extreme observations from the dataset if there is reason to think they must be fit with a different model, one the researcher will include in the analysis.

- *Missing at random:* Dropping outliers may also be accepted if the omitted data are missing at random as shown by the covariance matrix of interest to the researcher not being significantly different with or without the outliers. This would be rare.

- *True non-normal distribution*: If none of the foregoing situations explain outliers, then for a true non-normal distribution with extreme values, the researcher may transform the data to pull in outliers and proceed with the analysis without dropping outliers.

Simple outliers

Simple outliers are cases with extreme values with respect to a single variable. It is common to define outliers as cases which are more than plus or minus three standard deviations from the mean of the variable. Boxplots, discussed above, are a common means of identifying simple outliers.

Multivariate outliers

Multivariate outliers are cases with extreme values with respect to multiple variables. Multivariate outliers are operationally defined as cases which have a Cook's Distance greater than some cutoff (some use a cutoff of 1; some use 4/[n - p], where p is the number of parameters in the model; some use 4/[n - k - 1], where n is the number of cases and k is the number of independents.) Leverage is another related way of defining multivariate outliers, with outliers defined as having a leverage value greater than some cutoff (some use .5; others use 2p/n, where p is the number of parameters including the intercept). Mahalanobis distance is a third and very common measure for multivariate outliers. Cases with

the highest Mahalanobis D-square values are the most likely candidates to be considered outliers and should be examined.

These measures are discussed in the separate "blue book" volume on multiple regression. In SPSS, select Analyze, Regression, Linear; click the Save button; check Cook's, Mahalanobis, and/or leverage values.

Winsorizing data

An alternative to dropping outliers is to winsorize them. By arbitrarily reducing the values of outliers, winsorizing may bring data into an acceptably normal distribution, though before winsorizing, most researchers would try various data transforms to get acceptable skew and kurtosis. Bollinger & Chandra (2005) and others have found that in some circumstances, by pulling observations toward the mean, it is possible to introduce bias.

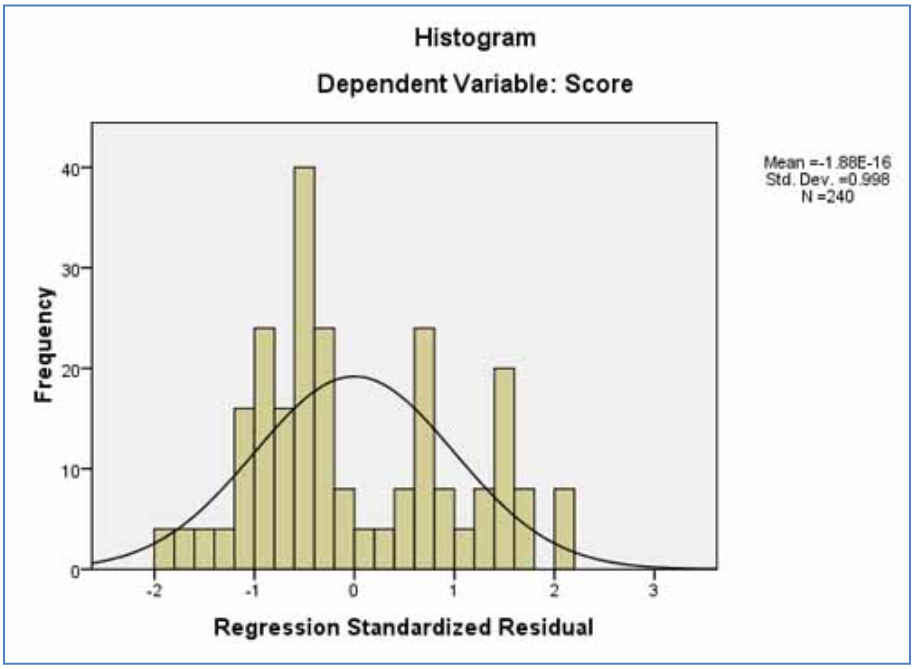
In winsorization, the researcher identifies the most extreme high and low values deemed reliable, then resets any more extreme cases to these clamping values. Rather than use absolute values, the clamping values may be set in percentile terms, such as the 5th and 95th percentile values. All more extreme values are set to the values of the 5th or 95th percentile case. Stata provides the commands winsor and wincorr to winsorize data or correlate data on a winsorized basis. SAS macros exist for winsorizing. In SPSS, winsorizing can be accomplished through the Recode procedure.

NORMALLY DISTRIBUTED ERROR

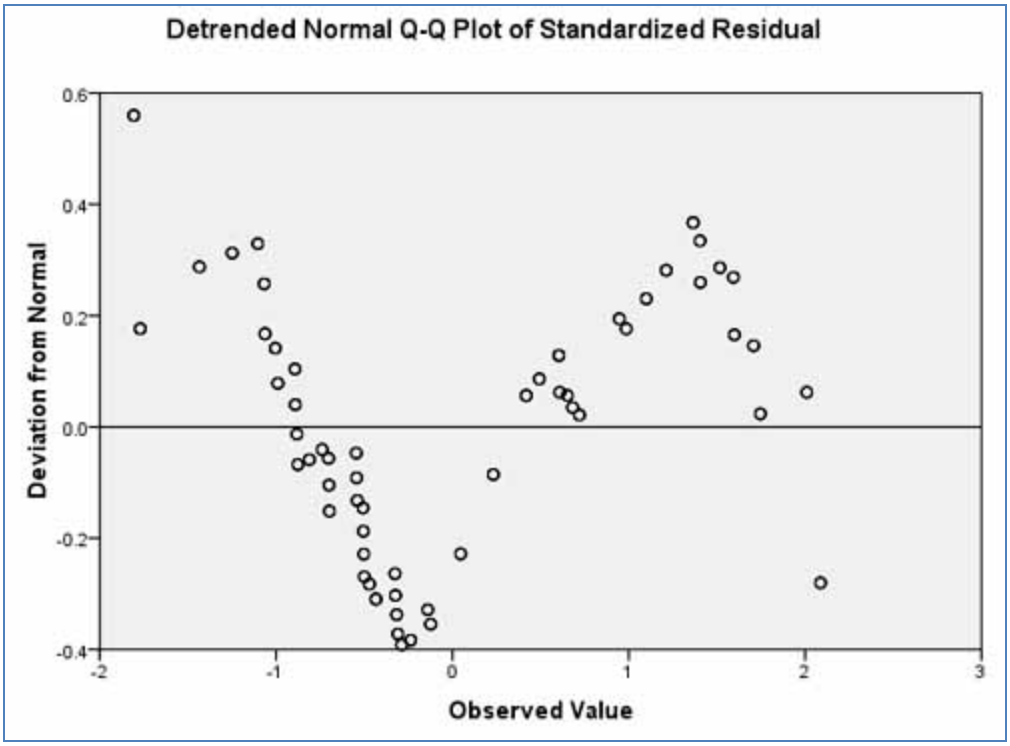
Histogram of standardized residuals

The histogram of standardized residuals should show a roughly normal curve when the assumption of regression and most other techniques is met that error terms are normally distributed. In any predictive technique, the expectation is normal distribution of error, with the largest number of predictions being at or near zero and then trailing off into "high prediction" and "low prediction" tails. SPSS output shows the normal curve superimposed on such histograms. In SPSS, select Graphs, Histogram. Or select Analyze, Regression, Linear; click Plots; check Histogram.

In the example below, Score is predicted from Seniority. In this example, graphical inspection leaves doubt whether error is normally distributed.

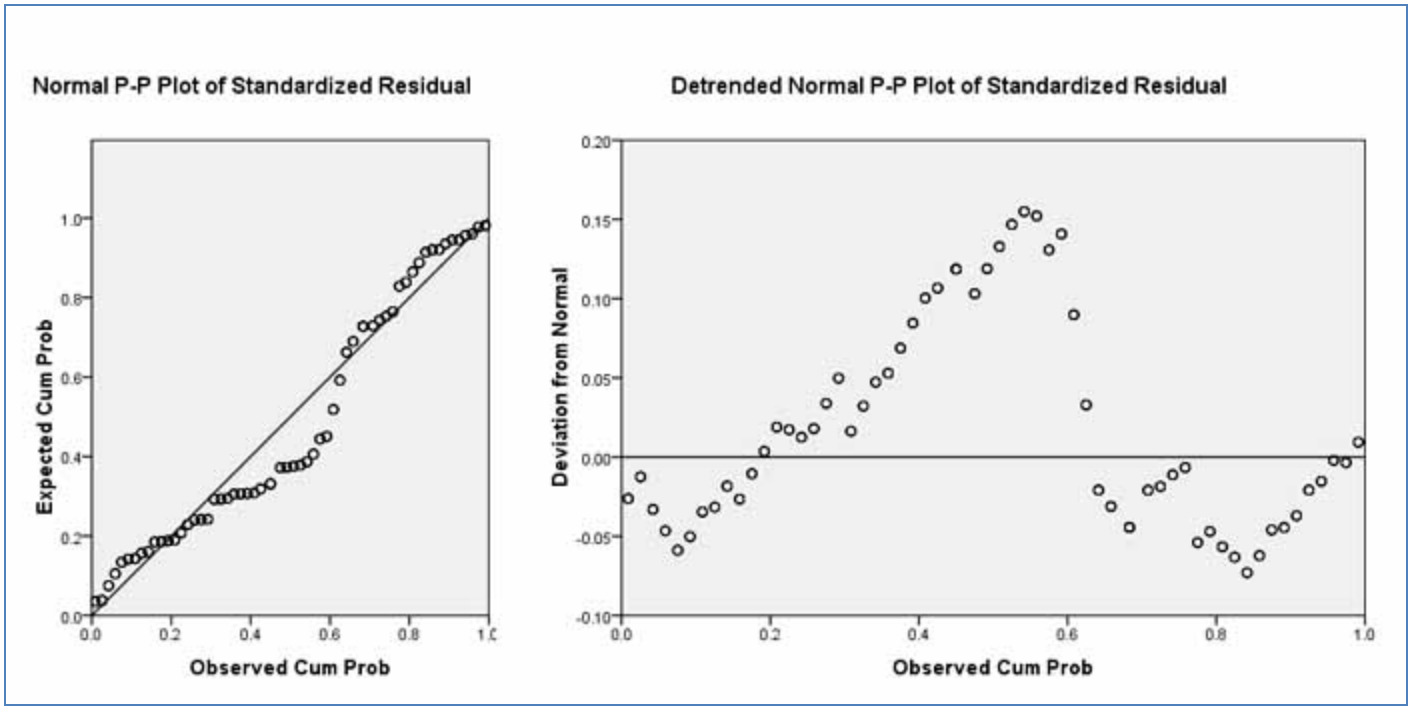


When the residuals for the example above are saved, the SPSS Analyze, Descriptive Statistics, Explore menu choices will generate, among other statistics, the skew and kurtosis, which are .436 and -.840 respectively for these data - within normal bounds. However, as the detrended Q-Q plot for these residuals shows (below), residuals are below normal expectations for middle ranges and above for high and low ranges. This is indicative of a bimodal rather than normal distribution of error. It is also an example of where the skew and kurtosis "rules of thumb" give misleading average values.



A normal probability plot

The normal probability plot, also called a P-P Plot, is an alternative method, plotting observed cumulative probabilities of occurrence of the standardized residuals on the Y axis and of expected normal probabilities of occurrence on the X axis, such that a 45-degree line will appear when the observed conforms to the normally expected and the assumption of normally distributed error is met. In the detrended version, the pattern of points above and below the horizontal 0 line should be random (unlike the illustration below). *P-P plots* are found in SPSS under Graphs, P-P plots. Or select Analyze, Descriptive Statistics, P-P plots. Or select Analyze, Regression, Linear; click Plots; check Normal probability plot.



Kolmogorov-Smirnov and other normality tests

In multiple regression one can use the Save option to save residuals to a variable which by default is labeled res_1. Other procedures likewise allow residuals to be saved. One can then use the SPSS Examine procedure to apply tests of normality, such as the Kolmogorov-Smirnov (Lilliefors) and Shapiro-Wilks normality tests, to this variable. Select Analyze, Descriptive Statistics, Examine; then under the Plots button, select "Normal probability plots with tests." These tests are discussed above, in the section on testing for normal distribution. Both tests should be non-significant if residuals are normally distributed.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	.163	240	.000	.940	240	.000

a. Lilliefors Significance Correction

HOMOGENEITY OF VARIANCES

Overview

Homogeneity of variances tests are required for analysis of variance (Anova) and certain other procedures.

Levene's test of homogeneity of variances

Levene's test of homogeneity of variance, which is the most common test, tests the assumption that each group (category) of one or more categorical independent variables has the same variance on an interval dependent. If the Levene statistic is significant at the .05 level or better, the researcher rejects the null hypothesis that the groups have equal variances. The Levene test is robust in the face of departures from normality. Levene's test appears in the SPSS procedures ONEWAY and T-TEST as well as EXAMINE. Usually it is based on deviations from the group mean. However, the EXAMINE module's spread-level plot option also prints versions for Levene's test based on deviations from the group median, deviations from the group median adjusted for degrees of freedom, and the 5% trimmed group mean. Levene's test is more robust in the face of non-normality than more traditional tests like Bartlett's test. In SPSS, select Analyze, Compare Means, One-Way ANOVA; click the Options buttion; check Homogeneity of variance test. Example.

Brown & Forsythe's test of homogeneity of variances

This test is based on criticisms of the Levene test. It tests for equality of group means. The Brown-Forsythe test is more robust than the Levine test when groups are unequal in size and the absolute deviation scores (deviations from the group means) are highly skewed, causing a violation of the normality assumption and the assumption of equal variances. In SPSS, select Analyze, Compare Means, One-Way ANOVA; click Options; select Brown-Forsythe.

Example

In the example below, Zodiac (Zodiac sign) is used to predict Polviews (liberal or conservative). As expected, the ANOVA is non-significant, indicating the Zodiac does not predict Polviews. Because the Levene statistic is not significant, the researcher fails to reject the null hypothesis that the groups have equal variances. Frequencies for Zodiac, not shown here, show group sizes not be be markedly different, so the results of the Levene test are accepted. However, were the group sizes markedly different, the Brown & Forsyth test would be used. For these data,

the Brown & Forsyth test is also non-significant and thus not different in inference from Levene's test.

ANOVA

Think of Self as Liberal or Conservative

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	22.591	11	2.054	1.120	.341
Within Groups	2601.119	1419	1.833		
Total	2623.711	1430			

Test of Homogeneity of Variances

Think of Self as Liberal or Conservative

Levene Statistic	df1	df2	Sig.
.579	11	1419	.847

Robust Tests of Equality of Means

Think of Self as Liberal or Conservative

	Statistic[a]	df1	df2	Sig.
Welch	1.127	11	547.334	.337
Brown-Forsythe	1.112	11	1358.122	.348

a. Asymptotically F distributed.

Welch test

The Welch test is an alternative to the ANOVA F test and is used when equality of group means cannot be assumed, as it cannot in this example. However, for this example, the Welch test also finds Zodiac not related to Polviews at a significant level.

Bartlett's test of homogeneity of variance

This is an older, alternative test. Bartlett's test is a chi-square statistic with (k-1) degrees of freedom, where k is the number of categories in the independent variable. The Bartlett's test is dependent on meeting the assumption of normality and therefore Levene's test has now largely replaced it.

F-max test

This test arises in the context of ANOVA. If the ratio of the largest to smallest size group in ANOVA is not very unequal (ex., is 4:1 or less), and if F-max (the ratio of the variance in the largest group to the variance in the smallest group) is 10:1 or less, then a rule of thumb is that homogeneity of variances is assumed not to be a problem. If groups are more unequal than 4:1, then the ratio of variance may need to be as low as 3:1.

SPHERICITY

Purpose. In a repeated measures design, the univariate ANOVA tables will not be interpreted properly unless the variance/covariance matrix of the dependent variables is circular in form. Sphericity is an assumption of repeated measures MANOVA, for instance. To conclude that sphericity is not violated, *Bartlett's test of sphericity* should not be significant. In SPSS, choose Analyze, General Linear Model, Multivariate; click the Options button; check Residual SSCP matrix.

HOMOGENEITY OF VARIANCE-COVARIANCE MATRICES

Box's M test

Box'x M tests the multivariate homogeneity of variances and covariances, as required by MANOVA and some other procedures. When M is not significant, the researcher accepts the null hypothesis that groups do not differ. It has been shown to be a conservative test, failing to reject the null hypothesis too often. It also is highly sensitive to violations of multivariate normality. In SPSS, choose Analyze, General Linear Model, Multivariate; click the Options button; check Homogeneity tests.

HOMOGENEITY OF REGRESSIONS / TEST OF PARALLELISM

Analysis of variance

In Ancova and Mancova, the slopes of the regression lines should be the same for each group formed by the categorical variables and measured on the dependents. The more this assumption is violated, the more conservative ANCOVA and

Mancova become (the more likely to make Type II errors - failing to reject null hypotheses, resulting in false negatives).

Violation of the homogeneity of regressions assumption indicates an interaction effect between the covariate(s) and the factor(s). When running an ANCOVA or MANCOVA model in SPSS, include in the model options the interactions between the covariate(s) and each independent factor -- any significant interaction effects indicate that the assumption of homogeneity of regression coefficients has been violated.

Homogeneity of regression

Homogeneity of regression in SPSS can be tested under the Model button of Analyze, General Linear Model, Univariate; select Custom under the Model button; enter a model with all main effects of the factors and covariates and the interaction of the covariate(s) with the factor(s). These interaction effects should be non-significant if the homogeneity of regressions assumption is met.

Parallelism tests

Tests of parallelism are similar in certain procedures, assuring that the slope of regression lines is the same for each level of a categorical grouping variable. Ordinal regression and the probit response model, for instance, come with a test of parallelism in SPSS and other packages.

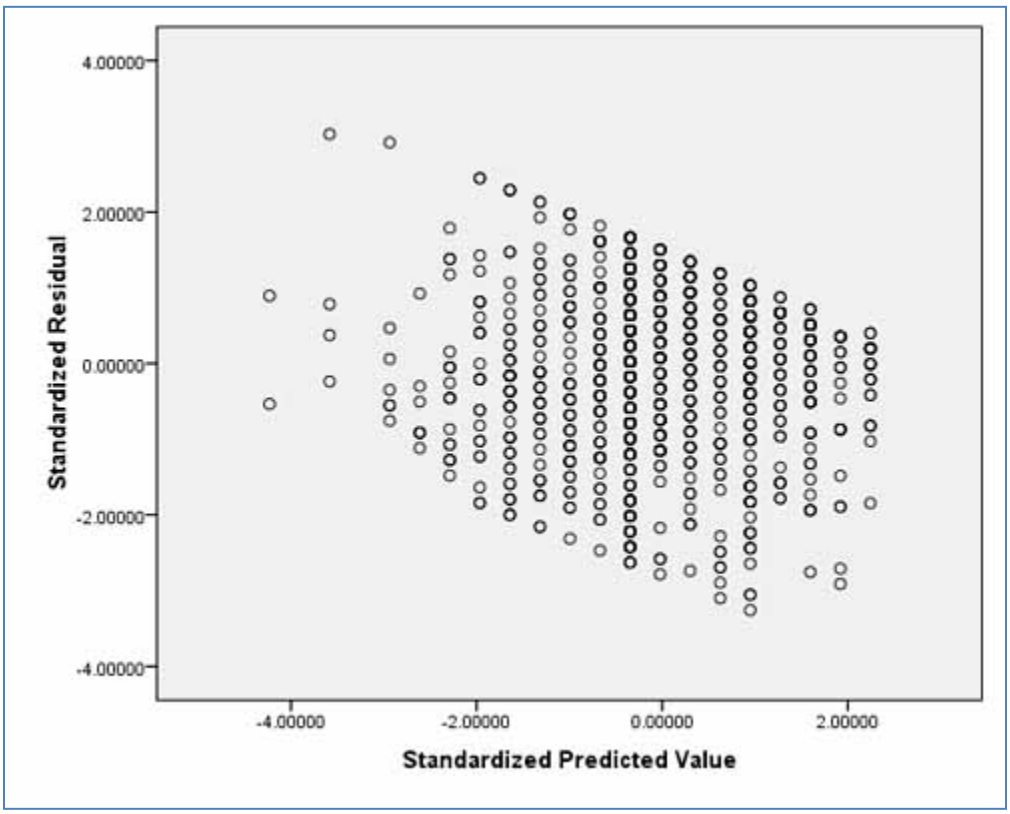
HOMOSCEDASTICITY

Graphical method

Homoscedasticity means the relationship under investigation is the same for the entire range of the dependent variable. Lack of homoscedasticity is shown by higher errors (residuals) for some portions of the range compared to others. When the homoscedasticity assumption is met, residuals will form a patternless cloud of dots. Lack of homoscedasticity is most easily seen in a standardized scatterplot. This scatterplot of the standardized predicted dependent variable (ZPR_1 in SPSS) against the standardized residuals (ZRE_1 in SPSS, or one may use studentized deleted residuals, SRE_1; or one might use the observed dependent vs. ZRE_1 or SRE_1) should show a random pattern across the entire range of ZPR_1 when, in regression, error is homoscedastic -- that is, when the regression

model is equally accurate across the range of the dependent. Sometimes a plot of residuals by the actual dependent values is used to test homoscedasticity also. In SPSS, select Analyze, Regression, Linear; click Plots.

If ZPR_1, ZRE_1, or other needed variables have been saved, you can also use Graphs, Legacy Dialogs, Scatter/Dot. In the output below, for instance, education was used to predict income and the standardized predicted and residual values were saved. The plot is largely a cloud (indicating homoscedasticity) but there is some pattern showing that higher predicted values have lower residuals (lack of homoscedasticity).



Weighted least squares regression

WLS is a commonly used strategy for dealing with heteroscedasticity in a linear regression context. See the separate "blue book" volume on WLS.

Goldfeld-Quandt test

This test is a formal test of homoscedasticity used when errors take a funnel (fan) shape. Though not directly supported by SPSS, this test involves running separate

regressions on the upper and lower observations in terms of values on the dependent, then conducting an F-test on the ratio of the error sum of squares (ESS) for the two regressions. To the extent that the ratio of the two ESS coefficients approaches 1.0, the homoscedasticity assumption is met.

Glejser test

This test is a formal test of homoscedasticity used when errors take a bow-tie (bimodal) shape and when sample size is not small. The residuals are regressed on the independent variable (or some function of it) and if the independent has a significant b coefficient, then the researcher concludes that homoscedasticity cannot be assumed.

Park test

The Park test is a related test of homoscedasticity. The squared residuals (or their log) is regressed on the independent variable (or its log), and if the independent has a significant b coefficient, then the researcher concludes that homoscedasticity cannot be assumed.

Breusch-Pagan-Godfrey test

This test is another large-sample test of homoscedasticity. The squared residuals are standardized by dividing by the mean squared residual(regression sum of squares (RSS) divided by N), giving the generalized residuals. The generalized residuals are then regressed on all independent variables (m variables) suspected of causing heteroscedasticity. The error sum of squares (ESS) divided by 2, for this regression, follows a chi-square distribution with (m - 1) degrees of freedom. A finding of significance means the null hypothesis is rejected and homoscedasticity cannot be assumed. The B-P-G test is more powerful than Goldfeld-Quandt or White's tests.

White's test

White's test involves regressing the squared residuals on all independent variables plus all squares of independent variables plus all crossproducts of independent variables (for a total of t predictor variables in all). The resulting R^2 is multiplied by N, giving a statistic with a chi-square distribution with t degrees of freedom. If the null hypothesis is rejected, homoscedasticity cannot be assumed.

White's test does not require prior knowledge of the form of heteroscedasticity. It is less powerful than other tests such as Goldfeld-Quandt.

LINEARITY

Testing for nonlinearity is necessary because correlation, regression, and other members of the general linear model (GLM) assume linearity. A variety of methods are available.

Graphical methods

Simple inspection of scatterplots is a common if non-statistical method of determining if nonlinearity exists in a relationship. Even better, a plot of standardized residuals against standardized estimates (fitted values) of the dependent variable should show a random pattern when nonlinearity is absent. In regression, as a rule of thumb, an indicator of possible nonlinearity is when the standard deviation of the residuals exceeds the standard deviation of the dependent. Adding to a model nonlinear terms such as squares or cubes of an independent and then seeing if R^2 in regression or fit indexes in structural equation modeling improve significantly is another way of testing for nonlinearity.

Curve fitting with R-squared difference tests

The SPSS curve fitting module (Analyze, Regression, Curve Fit) will calculater R-squared for a linear model compared to a variety of nonlinear models. One can then employ an F test of R-squared difference between models to see if nonlinear model has a significantly higher R-squared than a linear one. Note that some nonlinear models cannot be fitted if there are missing values, so interpolating using the Missing Values module may be a necessary prerequisite if the researcher wishes to compare these models. Warning: this method involves multiple comparisons which make the nominal significance level more lenient than it appears, so a Bonferroni or other adjustment may be necessary depending on the number of comparisons made.

ANOVA test of linearity

One can compute an ANOVA table for the linear and nonlinear components of any pair of variables. If the F significance value for the nonlinear component is below the critical value (ex., < .05), then there is significant nonlinearity. In SPSS, select

Analyze, Compare Means, Means; click Options; check Linearity test. Likewise, the Contrasts option in ANOVA can be used as a test for the existence of linear, quadratic, and other polynomial relationships.

Eta, the correlation ratio

Eta is a coefficient of nonlinear association. For linear relationships, eta equals the correlation coefficient (Pearson's r). For nonlinear relationships it is greater -- hence the difference between eta and r is a measure of the extent of nonlinearity of relationship. Eta is discussed in the separate "blue book" volume on measures of association. In SPSS, select Analyze, Compare Means, Means; click Options; check Anova table and eta.

Adding nonlinear terms to a model

A square term corresponds to a nonlinear pattern with a single curve, upward or downward. A cubic term corresponds to a nonlinear up-then-down or down-then-up pattern. In general, the number of curves is one less than the order of the polynomial term -- for instance, the 4th power will produce three curves. A more shotgun approach is to add all possible power terms to the model, then drop those which are non-significant.

Ramsey's RESET test (regression specification error test)

Ramsey's general test of specification error of functional form is an F test of differences of R^2 under linear versus nonlinear assumptions. It is commonly used in time series analysis to test whether power transforms need to be added to the model. For a linear model which is properly specified in functional form, nonlinear transforms of the fitted values should not be useful in predicting the dependent variable. While STATA and some packages label the RESET test as a test to see if there are "no omitted variables," it is a linearity test, not a general specification test. It tests if any nonlinear transforms of the specified independent variables have been omitted. It does not test whether other relevant linear or nonlinear variables have been omitted.

1. Run the regression to obtain R_o^2, the original multiple correlation.
2. Save the predicted values (Y's).
3. Re-run the regression using power functions of the predicted values (ex., their squares and cubes) as additional independents for the Ramsey RESET

test of functional form where testing that none of the independents is nonlinearly related to the dependent. Alternatively, re-run the regression using power functions of the independent variables to test them individually.

4. Obtain R_n^2, the new multiple correlation.
5. Apply the F test, where $F = (R_n^2 - R_o^2)/[(1 - R_n^2)/(n-p)]$, where n is sample size and p is the number of parameters in the new model.
6. Interpret F: For an adequately specified model, F should be non-significant.

MULTICOLLINEARITY

Multicollinearity is an unacceptably high level of intercorrelation among the independents, such that the effects of the independents cannot be separated. Under multicollinearity, estimates are unbiased but assessments of the relative strength of the explanatory variables and their joint effect are unreliable. (That is, beta weights and R-squares cannot be interpreted reliably even though predicted values are still the best estimate using the given independents). As a rule of thumb, intercorrelation among the independents above .80 signals a possible problem. Likewise, high multicollinearity is signalled when high R-squared and significant F tests of the model occur in combination with non-significant t-tests of coefficients.

Example

An example of multicollinearity occurred in a Bureau of Labor Statistics study of the price of camcorders. The initial model included the dummy variables Sony and 8mm, both of which corresponded to high price. However, since Sony was the only manufacturer of 8mm camcorders at the time, the Sony and 8 mm dummy variables were multicollinear. A similar multicollinearity occurred in a BLA study of washing machines, where it was found that "capacity" and "number of cycles" were multicollinear. In each study, one of the collinear variables had to be dropped from the model.

Whereas perfect multicollinearity leads to infinite standard errors and indeterminant coefficients, the more common situation of high multicollinearity leads to large standard errors, large confidence intervals, and diminished power (the chance of Type II errors is high - thinking you do not have a relationship when in fact one exists - failure to reject the null hypothesis that the coefficients are not

different from zero). R-square is high. The coefficients and their standard errors will be sensitive to changes in just a few observations. Methods of handling high multicollinearity are discussed in the "blue book" volume on multiple regression.

Tolerance

Tolerance is defined as 1 - R-squared, where R-squared is the multiple R of a given independent regressed on all other independent variables. If the tolerance value is less than some cutoff value, usually .20, the independent should be dropped from the analysis due to multicollinearity. This is better than just using simple r > .80 since tolerance looks at the independent variable in relation to all other independents and thus takes interaction effects into account as well as simple correlations. In SPSS, select Analyze, Regression, linear; click Statistics; check Collinearity diagnostics.

Variance inflation factor, VIF

Note, the variance-inflation factor, VIF, may be used in lieu of tolerance as VIF is simply the reciprocal of tolerance. The rule of thumb is that VIF > 4.0 when multicollinearity is a problem. Some authors use the more lenient cut-off of VIF >= 5 when multicollinearity is a problem. In SPSS, select Analyze, Regression, linear; click Statistics; check Collinearity diagnostics.

Condition indices.

Discussed more extensively in the "blue book" volume on multiple regression, condition indices over 15 indicate possible multicollinearity problems and over 30 indicate serious multicollinearity problems. In SPSS, select Analyze, Regression, linear; click Statistics; check Collinearity diagnostics.

Multicollinearity in Structural Equation Models (SEM)

Standardized regression weights: Since all the latent variables in a SEM model have been assigned a metric of 1, all the standardized regression weights should be within the range of plus or minus 1. When there is a multicollinearity problem, a weight close to 1 indicates the two variables are close to being identical. When these two nearly identical latent variables are then used as causes of a third latent variable, the SEM method will have difficulty computing separate regression weights for the two paths from the nearly-equal variables and the third

variable. As a result it may well come up with one standardized regression weight greater than +1 and one weight less than -1 for these two paths.

Standard errors of the unstandardized regression weights: Likewise, when there are two nearly identical latent variables, and these two are used as causes of a third latent variable, the difficulty in computing separate regression weights may well be reflected in much larger standard errors for these paths than for other paths in the model, reflecting high multicollinearity of the two nearly identical variables.

Covariances of the parameter estimates: Likewise, the same difficulty in computing separate regression weights may well be reflected in high covariances of the parameter estimates for these paths - estimates much higher than the covariances of parameter estimates for other paths in the model.

Variance estimates: Another effect of the same multicollinearity syndrome may be negative variance estimates. In the example above of two nearly-identical latent variables causing a third latent variable, the variance estimate of this third variable may be negative.

DATA INDEPENDENCE

Lack of independence

Independent observations are assumed by most statistical procedures, including multiple regression, logistic regression, and members of the general linear model (GLM) family. Lack of independence occurs in three broad classes of research situations.

- *Repeated measures data.* Before-after studies, panel studies, and paired comparison data measure the same subject at multiple times, where the subject's response at time t is correlated with the same subject's response at time t+1. Repeated measures modifications for GLM and some other statistical procedures are available.

- *Time series data.* Time series, such as budget data over time, is non-independent because of autocorrelation: data at time t is correlated with time t+1 (or other lags). Time series analysis adjusts for autocorrelation.

- *Hierarchical and grouped data.* In general, whenever data are grouped in some way there is the possibility they are non-independent because knowing the group for a case helps predict its value. Because data are apt to cluster by group, predictions will err when not taking grouping into account, leading to correlated error. Grouping may be by hierarchical layer (ex., census tract and city in multilevel sampling of individuals within tracts within cities), by observation time (ex., individual test scores grouped by year administered), and by any grouping factor (ex., religion). While non-independence may be addressed through repeated measures adaptations of conventional statistical procedures or by entering the grouping factor as a set of dummy variables, linear mixed modeling (a.k.a., hierarchical linear modeling or multilevel modeling) is the most rigorous approach to non-independence.

Intra-class correlation (ICC)

A now-conventional test for non-independence is to construct a null linear mixed model with the dependent as level 1 and the grouping factor as level 2, without other predictors at either level. If the ICC is significant, there is significant non-independence in the data. Computing ICC is discussed in the separate "blue book" volume on linear mixed modeling.

Durbin-Watson coefficient

Independence may also tested by the Durbin-Watson coefficient, which uses studentized residuals. The Durbin-Watson statistic should be between 1.5 and 2.5 for independent observations. In SPSS regression output, it is found in the "Model Summary" table. In SPSS, select Analyze, Regression, Linear; click Statistics; check Durbin-Watson.

Graphical method

Residuals may be plotted against the case identification number, when cases are ordered by any factor potentially causing non-independence. This factor may be time, a grouping factor, or simply interview/data collection order. There should be no pattern to this plot if observations are independent. In SPSS, select Graphs, Scatterplot.

RANDOMNESS

Runs Test

The runs test is used to test for randomness in a sample. Note that this is a necessary but not sufficient test for random sampling. A non-random availability sample of, say, students in a class, may be a very biased representation of all students in a university, yet within the class the order of sampling may be random and the sample may pass the runs test. On the other hand, if a purportedly random sample fails the runs test, this indicates that there are unusual, non-random periodicities in the order of the sample inconsistent with random sampling. In SPSS, select Analyze, Nonparametric Tests, Runs.

ADDITIVITY

Tukey's Test for nonadditivity

Tukey's test for nonadditivity tests whether a set of items are nonadditive. If the test returns a finding of significance, then items are not additive. This means that there are multiplicative interaction effects within the set of items and the overall effect is not a simple sum of the individual main effects. The Tukey's test also estimates the power to which items in a set would need to be raised in order to be additive. In SPSS, select Analyze, Scale, Reliability Analysis; click Statistics; check Tukey's test of additivity. This test is further discussed in the separate Statistical Associates "Blue Book" volume on scales and measures.

Transforms for additivity

If data do not test additive, sometimes this can be corrected by taking the log of y (the dependent) as a transform.

EQUALITY OF MEANS

Hotelling's T-square

T-square is a multivariate test for equality of means among items in a dataset. In SPSS, select Analyze, Scale, Reliability Analysis; click Statistics; check Hotelling's T-square.

Bibliography

Bollinger, Christopher R. & Chandra , Amitabh (2005). Iatrogenic specification error: A cautionary tale of cleaning data. *Journal of Labor Economics*, 23(2), 235-257.

Boneau, C. A. (1960). The effect of violation of assumptions underlying the t-test. *Psychological Bulletin*, 57: 49-64.

Cohen, Jacob (1969). *Statistical power analysis for the behavioral sciences*. NY: Academic Press.

Hutcheson, Graeme and Nick Sofroniou (1999). *The multivariate social scientist: Introductory statistics using generalized linear models*. Thousand Oaks, CA: Sage Publications. Chapter two covers data screening for assumptions under GLM.

Lipsey, Mark W. & Wilson, David B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage Publications.

Schumacker, Randall E. and Richard G. Lomax (2004). *A beginner's guide to structural equation modeling, Second edition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Shapiro, S. S. & Wilk. M. B. (1965). An analysis of variance test for normality. *Biometrika*, 52(3), 591-599.

Steenkamp, J.E.M., & van Trijp, H.C.M. (1991). The use of LISREL in validating marketing constructs. International Journal of Research in Marketing, 8: 283-299.

Vasu, E. S. (1979). Non-normality in regression analysis: Monte Carlo investigation under the condition of multicollinearity. Working papers in methodology. Number 9. Institute for Research in Social Science, University of North Carolina. Chapel Hill, North Carolina, USA.

Statistical Associates Publishing
Blue Book Series

Association, Measures of
Assumptions, Testing of
Canonical Correlation
Case Studies
Cluster Analysis
Content Analysis
Correlation
Correlation, Partial
Correspondence Analysis
Cox Regression
Creating Simulated Datasets
Crosstabulation
Curve Fitting & Nonlinear Regression
Data Levels
Delphi Method in Quantitative Research
Discriminant Function Analysis
Ethnographic Research
Evaluation Research
Event History Analysis
Factor Analysis
Focus Groups
Game Theory
Generalized Linear Models/Generalized Estimating Equations
GLM (Multivariate), MANOVA, and MANCOVA
GLM (Univariate), ANOVA, and ANCOVA
GLM Repeated Measures
Grounded Theory
Hierarchical Linear Modeling/Multilevel Analysis/Linear Mixed Models
Integrating Theory in Research Articles and Dissertations
Latent Class Analysis

Life Tables and Kaplan-Meier Survival Analysis
Literature Reviews
Logistic Regression
Log-linear Analysis
Longitudinal Analysis
Missing Values Analysis & Data Imputation
Multidimensional Scaling
Multiple Regression
Narrative Analysis
Network Analysis
Ordinal Regression
Parametric Survival Analysis
Partial Least Squares Regression
Participant Observation
Path Analysis
Power Analysis
Probability
Probit Regression and Response Models
Reliability Analysis
Resampling
Research Designs
Sampling
Scales and Standard Measures
Significance Testing
Structural Equation Modeling
Survey Research
Two-Stage Least Squares Regression
Validity
Variance Components Analysis
Weighted Least Squares Regression

Statistical Associates Publishing
http://www.statisticalassociates.com
sa.publishers@gmail.com