

Chapter 2

Basic Statistical Concepts: I. Correlation and Regression

Chance only favors invention for minds
which are prepared for discoveries by
patient study and persevering efforts.

—Louis Pasteur

Overview

- Standardized and unstandardized variables
- Bivariate correlation and regression
- Partial correlation
- Multiple correlation and regression
- Statistical tests
- Bootstrapping

Standardized and unstandardized variables

- A *standardized variable* has a mean of 0 and a standard deviation of 1.00
- Converting raw scores to z scores is the most common way to standardize a variable:

$$z = \frac{X - M}{SD}$$

- Standardized estimates are calculated with standardized variables, and they are interpreted the same way for all variables

Standardized and unstandardized variables

- The range of absolute values for standardized estimates is sometimes, but not always, 0 to 1.00
- In contrast, the range of *unstandardized estimates* is determined by the original metrics (scales) of the unstandardized variables
- Unstandardized estimates cannot generally be directly compared across different variables
- The covariance is an example of an unstandardized statistic:

$$COV_{XY} = r_{XY} SD_X SD_Y$$

Standardized and unstandardized variables

- Although researchers generally prefer standardized estimates, there is a strong preference in the SEM literature for unstandardized estimates
- One reason is that the most widely used estimation methods in SEM assume the analysis of unstandardized variables

Standardized and unstandardized variables

- Cases where standardized estimates may be inappropriate:
 1. Longitudinal measurement of variables that show increasing (or decreasing) variability over time
 2. When the original metrics of the variables are meaningful rather than arbitrary (e.g., survival in years)
 3. A structural equation model is analyzed across multiple samples that differ in their variabilities

Bivariate correlation and regression

- The Pearson correlation r estimates the degree of linear association between two continuous variables X and Y
- It is calculated with standardized variables:

$$r_{XY} = \frac{\sum z_X z_Y}{df}$$

$$df = N - 1$$

- r_{XY}^2 indicates the proportion of explained (shared) variance

Bivariate correlation and regression

- The theoretical range of r_{XY} is -1.00 to 1.00 , but its actual range can be narrowed closer to zero if the
 1. relation between X and Y is not linear
 2. variance of either X or Y is relatively narrow (restricted)
 3. shapes of the frequency distributions of X and Y are very different
 4. reliability of the scores on either X or Y is low
- Statistical tests of correlations generally assume that the residuals are independent, normally distributed, and have uniform variances (i.e., homoscedasticity)

Bivariate correlation and regression

- Some other types of bivariate correlations

1. Pearson correlations (i.e., special cases of r_{XY}):

Coefficient	Estimates the association between
Point-biserial (r_{pb})	a dichotomous variable and a continuous variable
Phi ($\hat{\phi}$)	two dichotomous variables
Spearman's rank order (rho)	two ordinal variables

Bivariate correlation and regression

- Some other types of bivariate correlations
 2. Non-Pearson correlations that estimate r_{XY} assuming both variables were continuous and normally distributed

Coefficient	Applied to
Biserial	a dichotomous variable and a continuous variable
Polyserial	a categorical variable with two or more levels and a continuous one
Tetrachoric	two dichotomous variables
Polychoric	two categorical variables each with two or more levels

Bivariate correlation and regression

- Estimation of the non-Pearson correlations just listed requires special software, such as PRELIS in LISREL
- The analysis in SEM of non-Pearson correlations from variables that are not continuous is discussed in chapter 7

Bivariate correlation and regression

- The linear regression equation for unstandardized variables is:

$$\hat{Y} = BX + A$$

$B = r_{XY} (SD_Y / SD_X)$ is the *unstandardized regression coefficient* (i.e., the slope)

$A = M_Y - B M_X$ is the *intercept (constant)*

Bivariate correlation and regression

- The value of B is the predicted difference on Y , given a difference of 1 point on X
- The value of A equals \hat{Y} , given $X = 0$
- The values of B and A in a particular sample are those that minimize the sum of the squared residuals, $\sum (Y - \hat{Y})^2$
- That is, they satisfy the *least squares criterion*

Bivariate correlation and regression

- Example:

$$\hat{Y} = 2.30 X + 10.00$$

A 1-point difference on X predicts the same on Y of 2.30 points

$$\hat{Y} = 10.00, \text{ given } X = 0$$

- Because B is an unstandardized estimate, its values are not limited to a particular range
- Consequently, a large numerical value of B does not necessarily mean that X is an “important” or “strong” predictor of Y

Bivariate correlation and regression

- The regression equation for standardized variables is

$$\hat{Z}_Y = r_{XY} Z_X$$

r_{XY} is the standardized regression coefficient (i.e., the Pearson correlation)

- The standardized regression coefficient indicates the expected difference on Y , given a difference on X of one full standard deviation
- Unlike that of B , the value of r_{XY} is unaffected by the scale of either X or Y

Partial correlation

- Partial correlation concerns the idea of spuriousness
- If the observed relation between X and Y is due to one or more common causes, their association is *spurious*
- The coefficient $r_{XY \cdot W}$ removes the effect of a third variable W from both X and Y and re-estimates their association:

$$r_{XY \cdot W} = \frac{r_{XY} - r_{XW} r_{YW}}{\sqrt{(1 - r_{XW}^2)(1 - r_{YW}^2)}}$$

Partial correlation

- If the value of r_{XY} is substantial but that of $r_{XY \cdot W}$ is about zero, then the observed association between X and Y “disappears” when controlling for W (i.e., it is spurious)
- SEM readily allows the representation and estimation of possible spurious associations

Multiple correlation and regression

- Assume two predictors X_1 and X_2 and a single criterion Y
- The multiple correlation $R_{Y.12}$ is just the Pearson correlation between observed and predicted criterion scores, $r_{Y\hat{Y}}$
- The theoretical range of $R_{Y.12}$ is 0 to 1.00
- Basically the same factors that can restrict the range of r_{XY} can also affect $R_{Y.12}$
- $R_{Y.12}^2$ is the proportion of total variance in Y explained by X_1 and X_2 together, correcting for their overlap

Multiple correlation and regression

- The unstandardized regression equation is

$$\hat{Y} = B_1 X_1 + B_2 X_2 + A$$

B_1 and B_2 are the unstandardized regression coefficients

$A = M_Y - B_1 M_1 - B_2 M_2$ is the intercept (i.e., the value of \hat{Y} , given $X_1 = X_2 = 0$)

- The values of B_1 , B_2 , and A satisfy the least squares criterion in a particular sample

Multiple correlation and regression

- Example:

$$\hat{Y} = 5.40 X_1 + 3.65 X_2 + 12.00$$

A difference of one point on X_1 predicts a difference on Y of 5.40 points, controlling for X_2

A difference of one point on X_2 predicts a difference on Y of 3.65 points, controlling for X_1

$$\hat{Y} = 12.00, \text{ given } X_1 = X_2 = 0$$

Multiple correlation and regression

- The regression equation for standardized variables is

$$\hat{z}_Y = \beta_1 z_1 + \beta_2 z_2$$

β_1 and β_2 are the standardized regression coefficients, or beta weights

- *Beta weights* indicate the expected difference on Y in standard deviation units controlling for the other predictors
- Values of beta weights can be directly compared across different variables

Multiple correlation and regression

- Example:

$$\hat{z}_Y = .60 z_1 + .35 z_2$$

The expected difference on Y is .60 standard deviations, given a difference on X_1 of one full standard deviation and controlling for X_2

The expected difference on Y is .35 standard deviations, given a difference on X_2 of one full standard deviation and controlling for X_1

Multiple correlation and regression

- Formulas for β_1 and β_2 :

$$\beta_1 = \frac{r_{Y1} - r_{Y2} r_{12}}{1 - r_{12}^2} \quad \text{and} \quad \beta_2 = \frac{r_{Y2} - r_{Y1} r_{12}}{1 - r_{12}^2}$$

- β_1 and β_2 are *not* correlations, and their absolute values can exceed 1.00 (if so, suppression is indicated)

Multiple correlation and regression

- The squared multiple correlation can be expressed as:

$$R_{Y.12}^2 = \beta_1 r_{Y1} + \beta_2 r_{Y2}$$

- If $r_{12} = 0$ (i.e., the predictors are unrelated), then

1. $\beta_1 = r_{Y1}$ and $\beta_2 = r_{Y2}$	}	That is, there is no correction for correlated predictors
2. $R_{Y.12}^2 = r_{Y1}^2 + r_{Y2}^2$		

Multiple correlation and regression

- *Specification error* in multiple regression refers to the problem of excluded (omitted) predictors
- An excluded predictor accounts for some unique proportion of total criterion variance, but is not included in the analysis
- Predictors are typically excluded because they are not measured by the researcher

Multiple correlation and regression

- Suppose that X_1 is the included (measured) predictor and X_2 is the excluded (unmeasured) predictor
- r_{Y1} is the standardized regression coefficient when just X_1 is in the equation
- β_1 is the standardized regression coefficient when both X_1 and X_2 are in the equation

Multiple correlation and regression

- The difference between r_{Y1} and β_1 depends on r_{12} , the correlation between the included and excluded predictors
- If $r_{12} = 0$, then $r_{Y1} = \beta_1$; otherwise, $r_{Y1} \neq \beta_1$
- As r_{12} increases, the difference between r_{Y1} and β_1 gets larger
- Implication:

r_{Y1} may not accurately reflect the “true” predictive power of the included predictor, X_1 , if it (X_1) has a substantial correlation with the excluded predictor, X_2

Multiple correlation and regression

- Example: X_1 is the included predictor, X_2 is the excluded predictor—compare the difference between $r_{Y1} = .40$ and β_1 as a function of r_{12} (Table 2.4):

				Regression with both predictors
Case	Predictor	β	$R_{Y \cdot 12}$	
1. $r_{12} = 0$	X_1	.40	.72	← $r_{Y1} - \beta_1 = 0$
	X_2	.60		
2. $r_{12} = .20$	X_1	.29	.66	← $r_{Y1} - \beta_1 = .11$
	X_2	.54		
3. $r_{12} = .40$	X_1	.19	.62	← $r_{Y1} - \beta_1 = .21$
	X_2	.52		

Note. For all cases, X_2 is considered the omitted variable; $r_{Y1} = .40$ and $r_{Y2} = .60$.

Multiple correlation and regression

- Perhaps the typical consequence of omitting a predictor that is correlated with an included predictor is overestimation of the unique predictive power of the latter
- However, it can happen that the predictive power of the included predictor winds up being overestimated or even that the sign of the estimate is incorrect
- Both cases just described indicate suppression
- See Mauro (1990) for more information about the problem of omitted predictors in multiple regression

Multiple correlation and regression

- A general definition of *suppression* is that it occurs when either
 1. the absolute value of a predictor's beta weight is greater than its bivariate correlation with the criterion (e.g., $\beta_1 = .60$, $r_{Y1} = .40$)
 2. the two have different signs (e.g., $\beta_1 = .10$, $r_{Y1} = -.30$)
- It happens because of correction for correlated predictors in the estimation of regression coefficients

Multiple correlation and regression

- Some types of suppression:

1. *Classical*: One predictor is uncorrelated with the criterion but receives a nonzero regression weight when controlling for another predictor

Example: Given $r_{Y1} = 0$, $r_{Y2} = .60$, and $r_{12} = .50$:

$$R_{Y.12} = .69, \beta_1 = -.40, \beta_2 = .80$$

Multiple correlation and regression

- Some types of suppression:

2. *Negative*: The predictors have positive correlations with the criterion and each other, but one receives a negative regression weight

Example: Given $r_{Y1} = .19$, $r_{Y2} = .49$, and $r_{12} = .70$:

$$R_{Y.12} = .54, \beta_1 = -.30, \beta_2 = .70$$

3. *Reciprocal*: Can occur when two predictors correlate positively with the criterion but negatively with each other

Multiple correlation and regression

- Suppression can be viewed from the perspective of specification error
- That is, omitting a key predictor (i.e., one of a pair of predictors involved in a suppression effect) can lead to misleading results for the included predictor
- Suppression can occur in SEM, too
- See Smith, Ager, and Williams (1992) for more information about suppression in regression
- Maasen and Bakker (2001) discuss the estimation of suppression effects in SEM

Multiple correlation and regression

- The role of multiple regression in SEM is actually quite limited
- It may be used for data screening or to analyze only specific kinds of path models
- It is not generally useful for estimating effects of latent variables that correspond to hypothetical constructs
- However, the ideas of correction for correlated predictors and specification error generalize directly to SEM
- See J. Cohen, P. Cohen, West, and Aiken (2003) for more information about multiple regression

Statistical tests

- There have been many problems with the use of statistical tests in the behavioral sciences
- Some highly respected behavioral scientists have even called for a ban on statistical tests in psychology research journals—see Nix and Barnette (1998)
- Their use has also been criticized across several other disciplines, including nursing, medicine, and wildlife management (e.g., Anderson, Burnham, & W. Thompson, 2000)
- W. Thompson (2001) made a list of 402 citations of works in different disciplines that question the indiscriminant use of statistical tests:

<http://biology.uark.edu/Coop/Courses/thompson5.html>

Statistical tests

- Outcomes of statistical tests— p values—are also widely misunderstood
- The correct interpretation of $p < .05$ is that the probability of the data assuming a true null hypothesis and random sampling is $< .05$

Statistical tests

- Some examples of common misinterpretations for the case $p < .05$:
 - ✓ The probability of the null hypothesis is $< .05$
 - ✓ The probability that the results are due to sampling error (chance) is $< .05$
 - ✓ The probability that the decision taken to reject the null hypothesis is a Type I error is $< .05$

Statistical tests

- Some examples of common misinterpretations for the case $p < .05$:
 - ✓ The probability of the alternative hypothesis is $> .95$
 - ✓ The probability of replication is $> .95$
- See Carver (1978), Kline (2004), and Oakes (1986) for more examples of common myths about p values

Statistical tests

- The failure to reject some null hypothesis is a meaningful outcome only if the power of the test is adequate
- Power is probability of rejecting the null hypothesis when there is a real effect in the population
- Power varies directly with the magnitude of the real population effect and the sample size

Statistical tests

- Other factors that affect power include the
 - ✓ level of statistical significance (e.g., .05 vs. .01)
 - ✓ directionality of the alternative hypothesis (i.e., a one-tailed vs. a two-tailed test)
 - ✓ whether the samples are independent or dependent (i.e., a between-subjects vs. a within-subjects design)
 - ✓ the particular test statistic used (e.g., parametric vs. nonparametric)
 - ✓ score reliability

Statistical tests

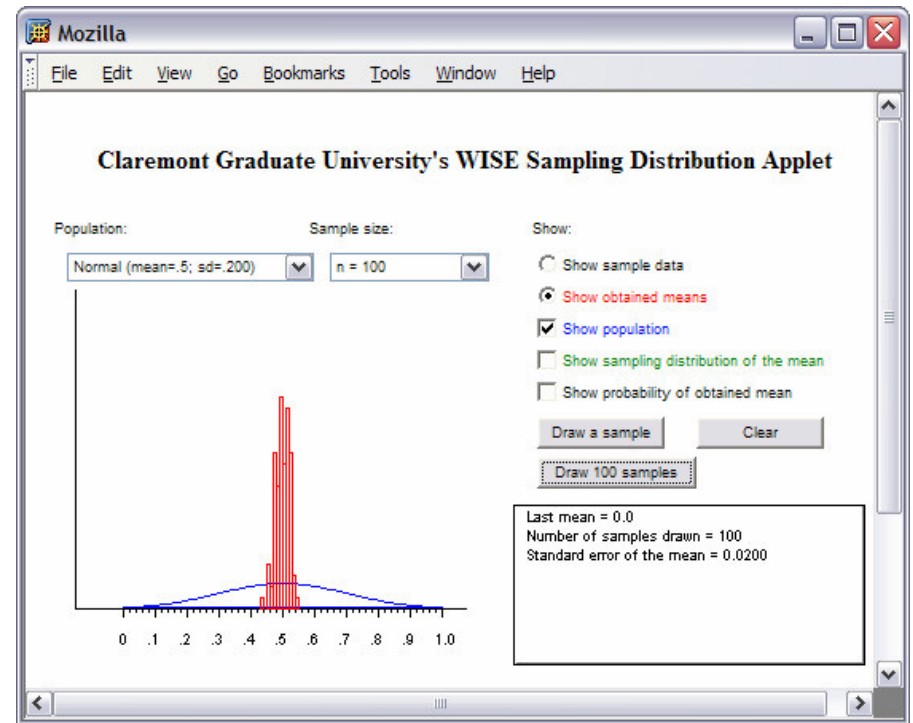
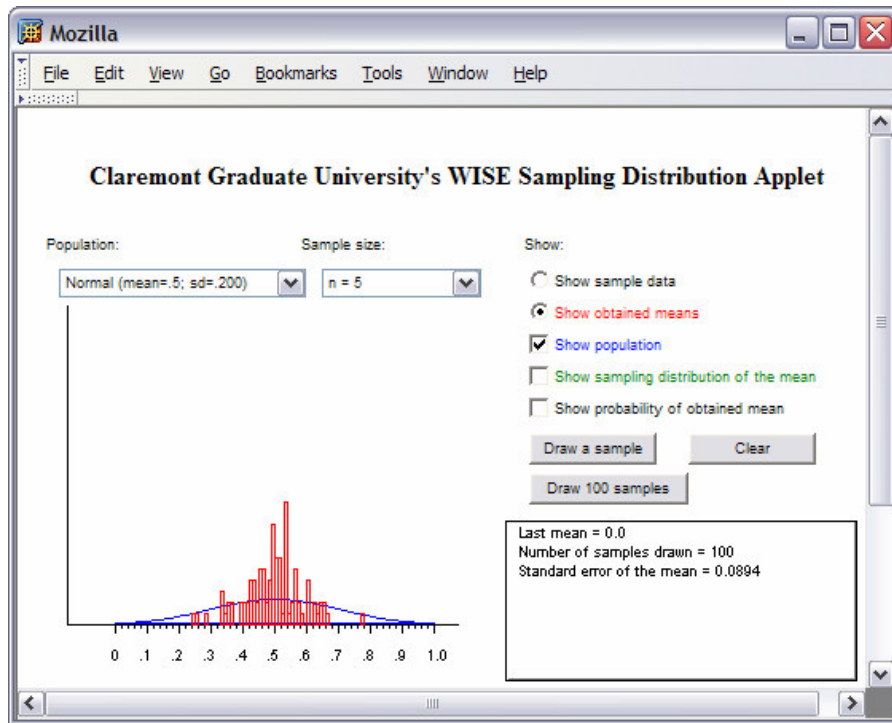
- This combination generally leads to the greatest power:
 - Large sample, the .05 level of statistical significance, a one-tailed test, a within-subjects design, a parametric test statistic, very reliable scores
- Power should be estimated when the study is planned but before the data are collected (e.g., Wilkinson and the Task Force on Statistical Inference, 1999)
- If power is not greater than .50, then tossing a coin has the same likelihood of being able to detect a real effect as conducting an empirical study
- There are methods for estimating power in SEM (chap. 6)

Statistical tests

- Perhaps the most basic form of a statistical test is the ratio of a sample statistic over its standard error
- A *standard error* is the standard deviation of a sampling distribution, which is a probability distribution of a statistic based on all possible random samples each based on the same number of cases
- Given constant variability among population cases, standard error varies inversely with sample size
- That is, distributions of statistics from larger samples are generally narrower than distributions of the same statistic from smaller samples

Statistical tests

- Claremont University WISE sampling distribution applet (<http://wise.cgu.edu/sdmmod/sdm.html>)
- Example: Sampling distributions for 100 samples for $N = 5$ (left) vs. $N = 100$ (right) given $\mu = .5$, $\sigma = .2$, and assuming normality:



Statistical tests

- There are “textbook” formulas for the standard errors of statistics with simple distributions
- By “simple” it is meant that the
 1. statistic estimates a single parameter
 2. shape of the distribution is not a function of the value of that parameter

Statistical tests

- There are some approximate methods for statistics with complex distributions that are amenable to hand calculation
- These methods generally assume large samples (i.e., they estimate asymptotic standard errors)
- But some statistics, such as R^2 , have distributions so complex that there is no approximate standard error formula amenable to hand calculation

Statistical tests

- Many standard errors in SEM—especially for effects of latent variables—are approximate
- This is one reason why one should not overinterpret results of statistical tests in SEM
- This is also greater emphasis on model-testing in SEM compared with traditional kinds of statistical analyses
- This perspective brings a higher-level perspective to the analysis

Bootstrapping

- This method is being used increasingly often in SEM
- Some SEM computer programs have built-in capabilities for bootstrapping
- Bootstrapping is a statistical resampling method developed by B. Efron in the late 1970s (e.g., Diaconis & Efron, 1983; Efron & Tibshirani, 1993)

Bootstrapping

- In *nonparametric bootstrapping*, cases from a raw data file are randomly selected with replacement to generate other data sets, usually with the same number of cases as the original
- Because of sampling with replacement, the
 1. same case can appear more than once in a generated data set
 2. composition of cases will vary somewhat across the generated samples

Bootstrapping

- When nonparametric bootstrapping is repeated many times, it simulates random sampling from a population
- Standard errors are estimated in this method as the standard deviation in the empirical sampling distribution of the same estimator across all generated samples

Bootstrapping

- Example: Here are raw scores and descriptive statistics for two groups:

	Group	
	1	2
	25	24
	26	25
	31	31
	29	28
	30	26
	30	26
	31	33
	32	28
	33	30
	33	29
M	30.00	28.00
s^2	7.33	8.00

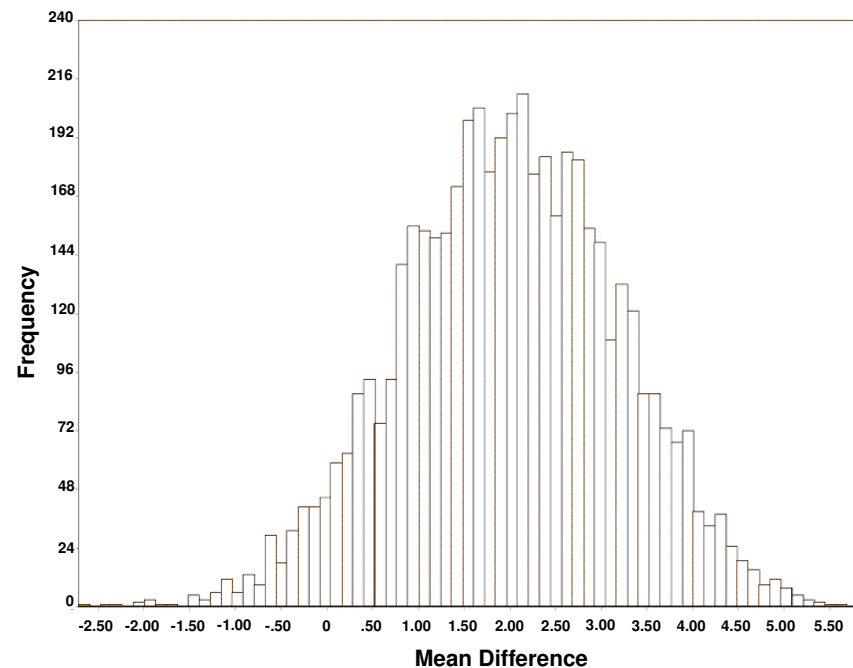
Bootstrapping

- The Bootstrap module of SimStat was used to generate 5,000 samples generated at random from the data set just shown
- The value of $M_1 - M_2$ from each generated sample was recorded
- The URL for SimStat:

<http://www.simstat.com/>

Bootstrapping

- Screenshot of the empirical sampling distribution of $M_1 - M_2$ across the 5,000 generated samples:



- The standard deviation in the above distribution is 1.195, which is a bootstrapped estimate of the standard error of $M_1 - M_2$

Bootstrapping

- There is actually little point in using nonparametric bootstrapping as just demonstrated to estimate standard errors for statistics with simple distributions, such as mean differences
- It is much more useful for statistics with complex distributions where estimation of standard errors by hand with simple formulas is not feasible
- This includes regressions coefficients for latent variables

Bootstrapping

- In *parametric bootstrapping*, the computer draws random samples from a probability density function with parameters specified by the researcher, not from a raw data file
- This is a kind of Monte Carlo method that is used in computer simulation studies of the properties of particular estimators, including those of many used in SEM that measure the fit of models to the data

Bootstrapping

- Bootstrapping is not a “magical” technique that can somehow compensate for
 - ✓ small or unrepresentative samples
 - ✓ distributions that are severely nonnormal
 - ✓ the absence of independent samples for replication
- In fact, bootstrapping can potentially magnify the effects of unusual features in a data set (Rodgers, 1999)

Bootstrapping

- Bootstrapping is used in SEM to estimate standard errors of model parameter estimates or fit indexes, which tend to have complex distributions
- There are also bootstrap methods to evaluate whole structural equation models (e.g., Bollen & Stine, 1993; Yung & Bentler, 1996) that are not discussed in the book

References

- Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null hypothesis testing: Problems, prevalence, and an alternative. *Journal of Wildlife Management*, 64, 912-923.
- Bollen, K. A., & Stine, R. A. (1993). Bootstrapping goodness-of-fit measures in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 111-135). Newbury Park, CA: Sage Publications.
- Carver, R. P. (1978). The case against significance testing. *Harvard Educational Review*, 48, 378-399.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahmah, NJ: Lawrence Erlbaum.
- Diaconis, P., & Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American*, 248, 116-130.
- Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman and Hall.
- Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, D.C.: American Psychological Association.
- Maasen, G. H., & Bakker, A. B. (2001). Suppressor variables in path models: Definitions and interpretations. *Sociological Methods and Research*, 30, 241-270.
- Mauro, R. (1990). Understanding L.O.V.E. (left-out variables error): A method for estimating the effects of omitted variables. *Psychological Bulletin*, 108, 314-329.
- Nix, T. W., & Barnette, J. J. (1998). The data analysis dilemma: Ban or abandon. A review of null hypothesis significance testing. *Research in the Schools*, 5, 3-14.
- Oakes, M. (1986). *Statistical inference*. New York: Wiley.

- Rodgers, J. L. (1999). The bootstrap, the jackknife, and the randomization test: A sampling taxonomy. *Multivariate Behavioral Research*, 34, 441-456.
- Smith, R. L., Ager, J. W., & Williams, D. L. (1992). Suppressor variables in multiple regression/correlation. *Educational and Psychological Measurement*, 52, 17-29.
- Thompson, W. L. (2001). 402 citations questioning the indiscriminate use of null hypothesis significance tests in observational studies. Retrieved November 11, 2001, from <http://biology.uark.edu/Coop/Courses/thompson5.html>
- Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594-604.
- Yung, Y.-F., & Bentler, P. M. (1996). Bootstrapping techniques in analysis of mean and covariance structures. In G. A. Marcoulides and R. E. Schumacker (Eds.), *Advanced structural equation modeling* (pp. 195-226). Mahwah, NJ: Lawrence Erlbaum.