Regression Primer

Knowing about regression analysis will help you to learn about SEM. Although the techniques considered next analyze observed variables only, their basic principles make up a core part of SEM. This includes the dependence of the results on not only what is measured (the data), but also on what is not measured, or omitted relevant variables, a kind of specification error. Some advice: *Even if you think that you already know a lot about regression, you should nevertheless read this primer carefully.* This is because many readers tell me that they learned something new after hearing about the issues outlined here. Next I assume that standard deviations (*SD*) for continuous variables are calculated as the square root of the sample variance $s^2 = SS/df$, where *SS* refers to the sum of squared deviations from the mean and the overall degrees of freedom are $df = N - 1$. Standardized scores, or normal deviates, are calculated as $z = (X - M)/SD$ for a continuous variable *X*.

BIVARIATE REGRESSION

Presented in Table R.1 are scores on three continuous variables. Considered next is bivariate regression for variables *X* and *Y*, but later we deal with the multiple regression analysis that also includes variable *W*. The unstandardized bivariate regression equation for predicting *Y* from *X*—also called regressing *Y* on *X*—takes the form

$$\hat{Y} = B_X X + A_X \qquad \text{(R.1)}$$

where $\hat{Y}$ refers to predicted scores. Equation R.1 describes a straight line where B_X, the unstandardized regression coefficient for predictor *X*, is the slope of the line, and A_X is the constant or intercept term, or the value of $\hat{Y}$, if $X = 0$. For the data in Table R.1,

$$\hat{Y} = 2.479X + 61.054$$

which says that a 1-point increase in *X* predicts an increase in *Y* of 2.479 points and that $\hat{Y} = 61.054$, given $X = 0$. Exercise 1 asks you to calculate these coefficients for the data in Table R.1.

The predicted scores defined by Equation R.1 make up a composite, or a weighted linear combination of the predictor and the intercept. The values of B_X and A_X in Equation R.1 are generally estimated with the method of **ordinary least squares** (OLS) so that the **least squares criterion** is satisfied. The latter means that the sum of squared residuals, or $\Sigma(Y - \hat{Y})^2$, is as small as possible *in a particular sample*. Consequently, OLS estimation capitalizes on chance variation, which implies that values of B_X and A_X will vary over samples. As we will see later, capitalization on chance is a greater problem in smaller versus larger samples.

Coefficient B_X in Equation R.1 is related to the Pearson correlation r_{XY} and the standard deviations of *X* and *Y* as follows:

$$B_X = r_{XY}\left(\frac{SD_Y}{SD_X}\right) \qquad \text{(R.2)}$$

A formula for r_{XY} is presented later, but for now we can

TABLE R.1. Example Data Set for Bivariate Regression and Multiple Regression

Case	*X*	*W*	*Y*	Case	*X*	*W*	*Y*
A	16	48	100	K	18	50	102
B	14	47	92	L	19	51	115
C	16	45	88	M	16	52	92
D	12	45	95	N	16	52	102
E	18	46	98	O	22	50	104
F	18	46	101	P	12	51	85
G	13	47	97	Q	20	54	118
H	16	48	98	R	14	53	105
I	18	49	110	S	21	52	111
J	22	49	124	T	17	53	122

Note. M_X = 16.900, SD_X = 3.007; M_W = 49.400, SD_W = 2.817; M_Y = 102.950, SD_Y = 10.870; r_{XY} = .686, r_{XW} = .272, r_{WY} = .499.

see in Equation R.2 that B_X is just a rearrangement of the expression for the covariance between X and Y, or $\text{cov}_{XY} = r_{XY}SD_XSD_Y$. Thus, B_X corresponds to the covariance structure of Equation R.1. Because B_X reflects the original metrics of X and Y, its value will change if the scale of either variable is altered (e.g., X is measured in centimeters instead of inches). For the same reason, values of B_X are not limited to a particular range. For example, it may be possible to derive values of B_X such as –7.50 or 1,225.80, depending on the raw score metrics of X and Y. Consequently, a numerical value of B_X that appears "large" does not necessarily mean that X is an important or strong predictor of Y.

The intercept A_X of Equation R.1 is related to both B_X and the means of both variables:

$$A_X = M_Y - B_X M_X \tag{R.3}$$

The term A_X represents the mean structure of Equation R.1 because it conveys information about the means of both variables (and the regression coefficient) albeit with a single number. As stated, $\hat{Y} = A_X$ when $X = 0$, but sometimes scores of zero are impossible on certain predictors (e.g., there is no IQ score of zero in conventional standardized metrics for such scores). If so, scores on X may be **centered**, or converted to mean deviations $x = X - M_X$, before analyzing the data. (Scores on Y are not centered.) Once centered, $x = 0$ corresponds to a score that equals the mean in the original (uncentered) scores, or $X = M_X$. When regressing Y on x, the value of the intercept A_x equals $\hat{Y}$ when $x = 0$; that is, the intercept is the predicted score on Y when X takes its average value in the raw data. Although centering generally changes the value of the intercept ($A_X \neq A_x$), centering does *not* affect the value of the unstandardized regression coefficient ($B_X = B_x$). Exercise 2 asks you to prove this point for the data in Table R.1.

Regression residuals, or $Y - \hat{Y}$, sum to zero and are uncorrelated with the predictor, or

$$r_{X(Y-\hat{Y})} = 0 \tag{R.4}$$

The equality represented in Equation R.4 is required in order for the computer to calculate unique values of the regression coefficient and intercept in a particular sample. Conceptually, assuming independence of residuals and predictors, or the **regression rule** (Kenny & Milan, 2012), permits estimation of the explanatory power of the latter (e.g., B_X for X in Equation R.1) controlling for omitted (unmeasured) predictors. Bollen (1989) referred to this assumption as **pseudo-isolation** of the measured predictor X from all other unmeasured predictors of Y. This term describes the essence of statistical control where B_X is

estimated, assuming that X is unrelated to all possible unmeasured predictors of Y.

The predictor and criterion in bivariate regression are theoretically interchangeable; that is, it is possible to regress Y on X or to regress X on Y in two separate analyses. Regressing X on Y would make less sense if X were measured before Y or if X is known to cause Y. Otherwise, the roles of predictor and criterion are not fixed in regression. The unstandardized regression equation for regressing X on Y is

$$\hat{X} = B_Y Y + A_Y \tag{R.5}$$

where the regression coefficient and intercept in Equation R.5 are defined, respectively as follows:

$$B_X = r_{XY}\left(\frac{SD_Y}{SD_X}\right) \text{ and } A_Y = M_X - B_Y M_Y \tag{R.6}$$

The expression for B_Y is nothing more than a different rearrangement of the same covariance, or $\text{cov}_{XY} = r_{XY}SD_XSD_Y$, compared with the expression for B_X (see Equation R.2). For the data in Table R.1, the unstandardized regression equation for predicting X from Y is

$$\hat{X} = .190\,Y - 2.631$$

which says that a 1-point increase in Y predicts an increase in X of .190 points and that $\hat{X} = -2.631$, given $Y = 0$. Presented in Figure R.1 are the unstandardized equations for regressing Y on X and for regressing X on Y for the data in Table R.1. In general, the two possible unstandardized prediction equations in bivariate regression are not identical. This is because the Y-on-X equation minimizes residuals on Y, but the X-on-Y equation minimizes residuals on X.

The equation for regressing Y on X when both variables are standardized (i.e., their scores are normal deviates, z) is

$$\hat{z}_Y = r_{XY} z_X \tag{R.7}$$

where $\hat{z}_Y$ is the predicted standardized score on Y and the Pearson correlation r_{XY} is the standardized regression coefficient. There is no intercept or constant term in Equation R.7 because the means of standardized variables equal zero. (Variances of standardized variables are 1.0.) For the data in Table R.1, $r_{XY} = .686$. Given $z_X = 1.0$ and $r_{XY} = .686$, then $\hat{z}_Y = .686\ (1.0)$, or .686;

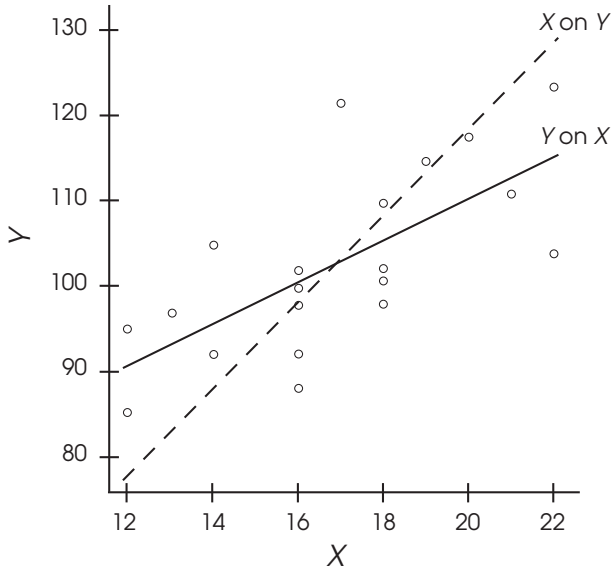


FIGURE R.1. Unstandardized prediction lines for regressing Y on X and for regressing X on Y for the data in Table R.1.

that is, a score one standard deviation above the mean on X predicts a score almost seven-tenths of a standard deviation above the mean on Y. A standardized regression coefficient thus equals the expected difference on Y in standard deviation units, given an increase on X of one full standard deviation. Unlike the unstandardized regression coefficient B_X (see Equation R.2), the value of the standardized regression coefficient (r_{XY}) is unaffected by the scale on either X or Y. It is true that (1) $r_{XY} = .686$ is also the standardized coefficient when regressing z_X on z_Y, and (2) the standardized prediction equation in this case is $\hat{z}_X = r_{XY}\, z_Y$.

There is a special relation between r_{XY} and the unstandardized predicted scores. If Y is regressed on X, for example, then

1. $r_{XY} = r_{Y\hat{Y}}$; that is, the bivariate correlation between X and Y equals the bivariate correlation between Y and $\hat{Y}$;
2. the observed variance in Y can be represented as the exact sum of the variances of the predicted scores and the residuals, or $s_Y^2 = s_{\hat{Y}}^2 + s_{Y-\hat{Y}}^2$; and
3. $r_{XY}^2 = s_{\hat{Y}}^2 / s_Y^2$, which says that the squared correlation between X and Y equals the ratio of the variance of the predicted scores over the variance of the observed scores on Y.

The equality just stated is the basis for interpreting squared correlations as proportions of explained variance, and a squared correlation is the **coefficient of determination**. For the data in Table R.1, $r_{XY}^2 = .686^2 = .470$, so we can say that X explains about 47.0% of the variance in Y, and vice versa. Exercise 3 asks you to verify the second and third equalities just described for the data in Table R.1.

When replication data are available, it is actually better to compare unstandardized regression coefficients, such as B_X, across different samples than to compare standardized regression coefficients, such as r_{XY}. This is especially true if those samples have different variances on X or Y. This is because the correlation r_{XY} is standardized based on the variability in a particular sample. If variances in a second sample are not the same, then the basis of standardization is not constant over the first and second samples. In contrast, the metric of B_X is that of the *raw scores* for variables X and Y, and these metrics are presumably constant over samples.

Unstandardized regression coefficients are also better when the scales of all variables are meaningful rather than arbitrary. Suppose that Y is the time to complete an athletic event and X is the number of hours spent in training. Assuming a negative covariance, the value of B_X would indicate the predicted decrease in performance time for every additional hour of training. In contrast, standardized coefficients describe the effect of training on performance in standard deviation units, which discard the original—and meaningful—scales of X and Y. The assumptions of bivariate regression are essentially the same as those of multiple regression. They are considered in the next section.

MULTIPLE REGRESSION

The logic of multiple regression is considered next for the case of two continuous predictors, X and W, and a continuous criterion Y, but the same ideas apply if there are three or more predictors. The form of the unstandardized equation for regressing Y on both X and W is

$$\hat{Y} = B_X X + B_W W + A_{X,W} \qquad \text{(R.8)}$$

where B_X and B_W are the **unstandardized partial regression coefficients** and $A_{X,W}$ is the intercept. The coefficient B_X estimates the change in Y, given a 1-point change in X while controlling for W. The coefficient B_W has the analogous meaning for the other predictor. The intercept $A_{X,W}$ equals the predicted score on Y when the scores on *both* predictors are zero, or $X = W = 0$. If zero is not a valid score on either predictor, then Y can be regressed on centered scores ($x = X - M_X$, $w = W - M_W$) instead of the original scores. If so, then $\hat{Y} = A_{x,w}$, given $X = M_X$ and $W = M_W$. As in bivariate regression, centering does not affect the values of the regression coefficients for each predictor in Equation R.8 (i.e., $B_X = B_x$, $B_W = B_w$).

The overall multiple correlation is actually just the Pearson correlation between the observed and predicted scores on the criterion, or $R_{Y\cdot X,W} = r_{Y\hat{Y}}$. Unlike bivariate correlations, though, the range of R is 0–1.0. The statistic R^2 equals the proportion of variance explained in Y by both predictors X and W, controlling for their intercorrelation. For the data in Table R.1, the unstandardized regression equation is

$$\hat{Y} = 2.147X + 1.302W + 2.340$$

and the multiple correlation equals .759. Given these results, we can say that

1. a 1-point change in X predicts an increase in Y of 2.147 points, controlling for W;
2. a 1-point change in W predicts an increase in Y of 1.302 points, controlling for X;
3. $\hat{Y} = 2.340$, given $X = W = 0$; and
4. the predictors explain $.759^2 = .576$, or about 57.6% of the total variance in Y, after taking account of their intercorrelation ($r_{XW} = .272$; Table R.1).

The regression equation just described defines a plane in three dimensions where the slope along the X-axis is 2.147, the slope along the W-axis is 1.302, and the Y-intercept for $X = W = 0$ is 2.340. This regression surface is plotted in Figure R.2 over the range of scores in Table R.1.

Equations for the unstandardized partial regression coefficients for each of two continuous predictors are

$$B_X = b_X\left(\frac{SD_Y}{SD_X}\right) \quad \text{and} \quad B_W = b_W\left(\frac{SD_W}{SD_Y}\right) \tag{R.9}$$

where b_X and b_W for X and W are, respectively, their **standardized partial regression coefficients**, also known as **beta weights**. Their formulas are listed next:

$$b_X = \frac{r_{XY} - r_{WY}r_{XW}}{1 - r_{XW}^2} \quad \text{and} \quad b_W = \frac{r_{WY} - r_{XY}r_{XW}}{1 - r_{XW}^2} \tag{R.10}$$

In the numerators of Equation R.10, the bivariate correlation of each predictor with the criterion is adjusted for the correlation of the other predictor with the criterion and for correlation between the two predictors. The denominators in Equation R.10 adjust the total standardized variance by removing the proportion shared by the two predictors. If the values of r_{XY}, r_{WY}, and r_{XW} vary over samples, then values of coefficients in Equations R.8–R.10 will also change.

Given three or more predictors, the formulas for the regression coefficients are more complicated but follow the same principles (see Cohen et al., 2003, pp. 636–642). If there is just a single predictor X, then $b_X = r_{XY}$. The intercept in Equation R.8 can be expressed as a function of the unstandardized partial regression coefficients and the means of all three variables as follows:

$$A_{X,W} = M_Y - B_X M_X - B_W M_W \tag{R.11}$$

The regression equation for standardized variables is

$$\hat{z}_Y = b_X\, r_{XY} + b_W\, r_{WY} \tag{R.12}$$

For the data in Table R.1, $b_X = .594$, which says that the difference on Y is expected to be about .60 stan-

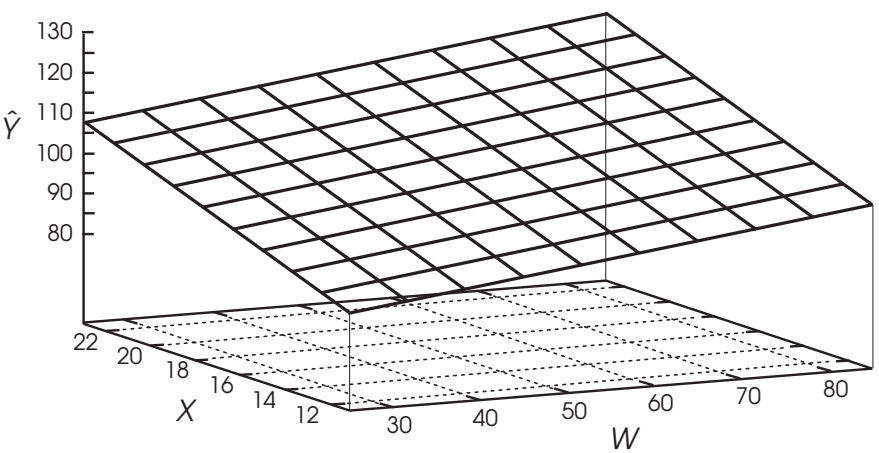


FIGURE R.2. Unstandardized regression surface for predicting Y from X and W for the data in Table R.1.

dard deviations large, given a difference on X of one full standard deviation, while we are controlling for W. The result $b_W = .337$ has the analogous meaning except that X is now statistically controlled. Because all variables have the same metric in the standardized solution, we can directly compare values of b_X with b_W and correctly infer that the relative predictive power of X is about 1.76 times that of W because the ratio .594/.337 = 1.76. In general, values of b can be directly compared across different predictors within the same sample, but unstandardized coefficients (B) are preferred for comparing results for the same predictor over different samples.

The statistic $R^2_{Y \cdot X, W}$ can also be expressed as a function of the beta weights and bivariate correlations of the predictors with the criterion. With two predictors,

$$R^2_{Y \cdot X, W} = b_X r_{XY} + b_W r_{WY} \qquad \text{(R.13)}$$

The role of beta weights as corrections for predictor overlap is also apparent in this equation. Specifically, if $r_{XW} = 0$ (the predictors are independent), then $b_X = r_{XY}$ and $b_W = r_{WY}$ (Equation R.10). This means that $R^2_{Y \cdot X, W}$ is just the sum of r^2_{XY} and r^2_{WY}. But if $r_{XW} \neq 0$ (the predictors covary), then b_X and b_W do not equal the corresponding bivariate correlations and $R^2_{Y \cdot X, W}$ is not the simple sum of r^2_{XY} and r^2_{WY} (it is less). Exercise 4 asks you to verify some of the facts about multiple regression just stated for the data in Table R.1.

Standard regression analyses do not require raw data files. This is because regression equations and values of R^2 can be calculated from summary statistics (e.g., Equation R.13), and many regression computer procedures read summary statistics as the input data. For example, the SPSS syntax listed next reads the summary statistics in Table R.1 and specifies the regression of Y on X and W. Four-decimal accuracy is recommended for matrix input:

```
comment table R.1, regress y on x, w.
matrix data variables=x w y/
contents=mean sd n corr
 /format=lower nodiagonal.
begin data
16.9000 49.4000 102.9500
3.0070 2.8172 10.8699
20 20 20
.2721
.6858 .4991
end data.
regression matrix=in(*)/
variables=x w y/
dependent=y
 /enter.
```

A drawback to conducting regression analyses with summaries statistics is that residuals cannot be calculated for individual cases.

Corrections for Bias

The statistic R^2 is a positively biased estimator of ρ^2 (rho-squared), the population proportion of explained variance. The degree of bias is greater in smaller samples or when the number of predictors is large relative to the number of cases. For example, if $N = 2$ in bivariate regression and there are no tied scores on X or Y, then r^2 must equal 1.0. Now suppose that $N = 100$ and $k = 99$, where k is the number of predictor variables. With so many predictors—in fact, the maximum number for $N = 100$—the value of R^2 must equal 1.0 because there can be no error variance with so many predictors, and this is true even for random numbers.

There are many corrections that downward adjust R^2 as a function of N and k. Perhaps the most familiar is Wherry's (1931) equation:

$$\hat{R}^2 = 1 - (1 - R^2)\left(\frac{N-1}{N-k-1}\right) \qquad \text{(R.14)}$$

where $\hat{R}^2$ is the **shrinkage-corrected estimate of ρ^2**. In small samples it can happen that $\hat{R}^2 < 0$; if so, then $\hat{R}^2$ is interpreted as though its value were zero. As the sample size increases for a constant number of predictors, values of $\hat{R}^2$ and R^2 are increasingly similar, and in very large samples they are essentially equal; that is, it is unnecessary to correct for positive bias in very large samples. Exercise 5 asks you to apply the Wherry correction to the data in Table R.1.

Assumptions

The statistical and conceptual assumptions of regression are strict, probably more so than many researchers realize. They are summarized next:

1. *Regression coefficients reflect unconditional linear relations only.* The estimate for B_X in Equation R.8 assumes that the linear relation between X and Y remains constant over all levels of (a) X itself, (b) the

other measured predictor, *W*, and (c) all unmeasured predictors. But if the relation between *X* and *Y* is appreciably curvilinear or conditional, the value of B_X could misrepresent predictive power. A conditional relation implies interaction, where the covariance between *X* and *Y* changes over the levels of at least one other predictor, measured or unmeasured. A curvilinear relation of *X* to *Y* is also conditional in the sense that the shape of the regression surface changes over the levels of *X* (e.g., Figure 7.7). How to represent curvilinear or interactive effects in regression analysis and SEM is considered in Chapter 7.

2. *All predictors are perfectly reliable (no measurement error).* This very strong assumption is necessary because there is no direct way in standard regression analysis to represent or control for less-than-perfect score reliability for the predictors. Consequences of minor violations of this requirement may not be critical, but more serious ones can result in substantial bias. This bias can affect not only the regression weights of predictors measured with error but also those of other predictors. It is difficult to anticipate the direction of this **propagation of measurement error**. Depending on sample intercorrelations, some absolute regression weights may be biased upward (too large), but others may be biased in the other direction (too small), or **attenuation bias**. There is no requirement that the criterion be measured without error, but the use of a psychometrically deficient measure of it can reduce the value of R^2. Note that measurement error in the criterion only affects the standardized regression coefficients, not the unstandardized ones. If the predictors are also measured with error, too, then these effects for the criterion could be amplified, diminished, or canceled out, but it is best not to hope for the absence of bias; see Williams et al. (2013) for more information about measurement error in regression analysis.

3. *Significance tests in regression assume that the residuals are normally distributed and homoscedastic*. The homoscedasticity assumption means that the residuals have constant variance across all levels of the predictors. Distributions of residuals can be heteroscedastic (the opposite of homoscedastic) or non-normal due to outliers, severe non-normality in the observed scores, more measurement error at some levels of the criterion or predictors, or a specification error. The residuals should always be inspected in regression analyses (see Cohen, Cohen, West, & Aiken, 2003, chap. 4). *Reports of regression analyses without comment on the residuals are inadequate*. Exercise 6 asks you to inspect the residuals for the multiple regression analysis of the data in Table R.1. Although there is no requirement in regression for normal distributions of the original scores, values of multiple correlations and absolute partial regression coefficients are reduced if the distributions for a predictor and the criterion have very different shapes, such as very positively skewed on one versus very negatively skewed on the other.

4. *There are no causal effects among the predictors (i.e., there is a single equation).* Because predictors and criteria are theoretically interchangeable in regression, such analyses can be viewed as strictly predictive. But sometimes the analysis is explicitly or implicitly motivated by causal hypotheses, where a researcher views the regression equation as a prototypical causal model with the predictors as causes and the criterion as their outcome (Cohen et al., 2003). If predictors in standard regression analyses are viewed as causal, then we must assume there are no causal effects among them. Specifically, standard regression analyses do not allow for indirect causal effects where one predictor, such as *X*, affects another, such as *W*, which in turn affects the criterion, *Y*. The indirect effect just described would be represented in SEM by the presumed causal order

$$X \longrightarrow W \longrightarrow Y$$

From a regression perspective, (1) variable *W* is both a predictor (of *Y*) and an outcome (of *X*), and (2) there are actually two equations, one for *W* another for *Y*. But standard regression techniques analyze a single equation at a time, in this case for just *Y*, and thus yield estimates of direct effects only. If there are appreciable indirect effects but such effects are not explicitly represented in the analysis,, then estimates of direct effects in standard regression analyses can be very wrong (Achen, 2005). The idea behind this type of bias is elaborated in Chapter 6, which concerns a graph-theoretic approach to causal inference.

5. *There is no specification error.* A few different kinds of potential mistakes involve specification error. These include the failure to estimate the correct functional form of relations between predictors and the criterion, such as assuming unconditional linear effects only when there are sizable curvilinear or interactive effects. Use of the incorrect estimation method is another kind of error. For example, OLS estimation is for continuous criteria, but dichotomous outcomes (e.g.,

pass–fail) generally require different methods, such as those used in logistic regression. Including predictors that are irrelevant in the population is a specification error. The concern is that an irrelevant predictor could in a particular sample relate to the criterion by sampling error alone, and this chance covariance may distort values of regression coefficients for other predictors. Omitting from the regression equation predictors that (1) account for some unique proportion of criterion variance and (2) covary with measured predictors is **left-out variables error**, described next.

LEFT-OUT VARIABLES ERROR

—or more lightheartedly described as the "heartbreak of L.O.V.E." (Mauro, 1990), this is a potentially serious specification error. As covariances between measured (included) and unmeasured (excluded) predictors increase, results based on the included predictors only tend to become progressively more biased. Suppose that $r_{XY} = .40$ and $r_{WY} = .60$ for, respectively, predictors X and W. A researcher measures only X and specifies it as the sole predictor of Y in a bivariate regression. In this analysis for the included predictor, the standardized regression coefficient is $r_{XY} = .40$. But if the researcher had the foresight to also measure W, the omitted predictor, and specify it along with X as predictors in a multiple regression analysis (e.g., Equation R.8), the beta weight for X in this analysis, b_X, may not equal .40. If not, then r_{XY} as a standardized regression coefficient with X as the sole predictor does not reflect the true relation of X to Y compared with b_X derived with both predictors in the equation.

The difference between r_{XY} and b_X varies with r_{XW}, the correlation between the included and omitted predictors. Specifically, if the included and omitted predictors are unrelated ($r_{XW} = 0$), there is no difference, or $r_{XY} = b_X = .40$ in this example because there is no correction for correlated predictors. Specifically, given

$$r_{XY} = .40,\ r_{WY} = .60, \text{ and } r_{XW} = 0$$

you can verify, using Equations R.10 and R.13, that the multiple regression results with both predictors are

$$b_X = .40,\ b_W = .60, \text{ and } R^2_{Y \cdot X,\, W} = .52$$

So we conclude that $r_{XY} = b_X = .40$ regardless of whether or not W is included in the regression equation, given $r_{XW} = 0$.

Now suppose that

$$r_{XY} = .40,\ r_{WY} = .60, \text{ and } r_{XW} = .60$$

Now we assume that the correlation between the included predictor X and the omitted predictor W is .60, not zero. In the bivariate analysis with X as the sole predictor, $r_{XY} = .40$ (the same as before), but now the results of the multiple regression analysis are

$$b_X = .06,\ b_W = .56, \text{ and } R^2_{Y \cdot X,\, W} = .36$$

Here the value of b_X is much lower than that of r_{XY}, respectively, .06 versus .40. This happens because coefficient b_X controls for $r_{XW} = .60$, whereas r_{XY} does not; thus, r_{XY} overestimates the relation between X and Y compared with b_X.

Omitting a predictor correlated with others in the equation does not always result in overestimation of the predictive power of an included predictor. For example, if X is the included predictor and W is the omitted predictor, it is also possible for the absolute value of r_{XY} in the bivariate analysis to be less than that of b_X when both predictors are included in the equation; that is, r_{XY} underestimates the relation indicated by b_X. It is also possible for r_{XY} and b_X to have different signs. Both cases just mentioned indicate suppression, described in more detail in the next section. But overestimation due to omission of a predictor may occur more often than underestimation (suppression). Also, the pattern of bias may be more complicated when there are several omitted variables (e.g., overestimation for some measured predictors, underestimation for others).

Predictors are typically excluded because they are not measured. This means that it is difficult to actually know by how much and in what direction(s) regression coefficients may be biased relative to what their values would be if all relevant predictors were included. But it is unrealistic to expect the researcher to know and be able to measure all relevant predictors. In this sense, all regression equations are probably misspecified to some degree. If omitted predictors are uncorrelated with included predictors, the consequences of left-out variables error may be slight; otherwise, the consequences may be more serious. Careful review of theory and research is the main way to avoid serious specification error by decreasing the potential number of left-out variables.

SUPPRESSION

Perhaps the most general definition is that suppression occurs when either (1) the absolute value of a predictor's beta weight is greater than that of its bivariate correlation with the criterion or (2) the two have different signs (see also Shieh, 2006). So defined, suppression implies that the estimated relation between a predictor and a criterion while controlling for other predictors is a "surprise," given the bivariate correlations. Suppose that X is the amount of psychotherapy, W is the degree of depression, and Y is the number of prior suicide attempts. The bivariate correlations in a hypothetical sample are

$$r_{XY} = .19,\ r_{WY} = .49, \text{ and } r_{XW} = .70$$

Based on these results, it might seem that psychotherapy is harmful because of its positive association with suicide attempts ($r_{XY} = .19$). When both predictors (psychotherapy and depression) are analyzed in multiple regression, however, the results are

$$b_X = -.30,\ b_W = .70, \text{ and } R^2_{Y \cdot X,\, W} = .29$$

The beta weight for psychotherapy (–.30) has the opposite sign of its bivariate correlation (.19), and the beta weight for depression (.70) exceeds its bivariate correlation (.49).

The results just described are due to controlling for other predictors. Here, people who are more depressed are more likely to be in psychotherapy ($r_{XW} = .70$) and also more likely to try to harm themselves ($r_{WY} = .49$). Correcting for these associations in multiple regression indicates that the relation of psychotherapy to suicide attempts is actually *negative* once depression is controlled. It is also true that the relation of depression to suicide is even *stronger* (here, more positive) once psychotherapy is controlled. Omit either psychotherapy or depression from the analysis—a left-out variables error—and the bivariate results with the remaining predictor are misleading.

The example just described concerns **negative suppression**, where the predictors have positive bivariate correlations with the criterion and with each other, but one receives a negative beta weight in the multiple regression analysis. A second type is **classical suppression**, where one predictor is uncorrelated with the criterion but receives a nonzero beta weight controlling for another predictor. For example, given the following correlations in a hypothetical sample,

$$r_{XY} = 0,\ r_{WY} = .60, \text{ and } r_{XW} = .50$$

the results of a multiple regression analysis are

$$b_X = -.40,\ b_W = .80, \text{ and } R^2_{Y \cdot X,\, W} = .48$$

This example of classical suppression (i.e., $r_{XY} = 0$, $b_X = -.40$) demonstrates that bivariate correlations of zero can mask true predictive relations once other variables are controlled. There is also **reciprocal suppression**, which can occur when two variables correlate positively with the criterion but negatively with each other. Some cases of suppression can be modeled in SEM as the result of inconsistent direct versus indirect effects of causally prior variables on outcome variables. These possibilities are explored later in the book.

PREDICTOR SELECTION AND ENTRY

An implication of suppression is that predictors should not be selected based on values of bivariate correlations with the criterion. These **zero-order associations** do not control for other predictors, so their values can be misleading compared with partial regression coefficients for the same variables. For the same reason, whether or not bivariate correlations with the criterion are statistically significant is also irrelevant concerning predictor selection. Although regression computer procedures make it easy to do so, researchers should avoid mindlessly dumping long lists of explanatory variables into regression equations in order to control for their effects (Achen, 2005). The risk is that even small but undetected nonlinearities or indirect effects among predictors can seriously bias partial regression coefficients. It is better to judiciously select the smallest number of predictors—those deemed essential based on extant theory or results of prior empirical studies.

Once selected, there are two basic ways to enter predictors into the equation: One is to enter all predictors at once, or **simultaneous (direct) entry**. The other is to enter them over a series of steps, or **sequential entry**. Entry order can be determined according to one of two different standards, theoretical (rational) or empirical (statistical). The rational standard corresponds to **hierarchical regression**, where you tell the computer a fixed order for entering the predictors. For example, sometimes demographic variables are entered at the first step, and then entered at the second step is a psychological variable of interest. This order not only controls for the demographic variables but also permits

evaluation of the predictive power of the psychological variable, over and beyond that of the simple demographic variables. The latter can be estimated as the increase in the squared multiple correlation, or ΔR^2, from that of step 1 with demographic predictors only to that of step 2 with all predictors in the equation.

An example of the statistical standard is **stepwise regression**, where the computer selects predictors for entry based solely on statistical significance; that is, which predictor, if entered into the equation, would have the smallest p value for the test of its partial regression coefficient? After selection, predictors at a later step can be removed from the equation according to p values (e.g., if $p \geq .05$ for a predictor in the equation at a particular step). The stepwise process stops when there could be no statistically significant ΔR^2 by adding more predictors. Variations on stepwise regression include **forward inclusion**, where selected predictors are not later removed from the equation, and **backward elimination**, which begins with all predictors in the equation and then automatically removes them, but such methods are directed by the computer, not you.

Problems of stepwise and related methods are so severe that they are actually banned in some journals (Thompson, 1995), and for good reasons, too. One problem is extreme capitalization on chance. Because every result in these methods is determined by p values in a particular sample, the findings are unlikely to replicate. Another problem is that not all stepwise regression procedures report p values that are corrected for the total number of variables that were considered for inclusion. Consequently, p values in stepwise computer output are generally too low, and absolute values of test statistics are too high; that is, the computer's choices could actually be wrong. Even worse, such methods give the false impression that the researcher does not have to think about predictor selection. Stepwise and related methods are anachronisms in modern data analysis. Said more plainly, death to stepwise regression, think for yourself (e.g., hierarchical entry)—see Whittingham, Stephens, Bradbury, and Freckleton (2006) for more information.

Once a final set of rationally selected predictors has been entered into the equation, they should *not* be subsequently removed if their regression coefficients are not statistically significant. To paraphrase Loehlin (2004), the researcher should *not* feel compelled to drop every predictor that is not significant. In smaller samples, the power of significance tests may be low, and removing a nonsignificant predictor can substantially alter the solution. If you had good reason for including a predictor, then it is better to leave it in the equation until replication indicates that the predictor does not appreciably relate to the criterion.

PARTIAL AND PART CORRELATION

The concept of partial correlation concerns the idea of **spuriousness**: If the observed relation between two variables is wholly due to one or more common cause(s), their association is spurious. Consider these bivariate correlations between vocabulary breadth (Y), foot length (X), and age (W) in a hypothetical sample of elementary school children:

$$r_{XY} = .50,\ r_{WY} = .60, \text{ and } r_{XW} = .80$$

Although the correlation between foot length X and vocabulary breadth Y is fairly substantial (.50), it is hardly surprising because both are caused by a third variable, age W (i.e., maturation).

The **first-order partial correlation** $r_{XY \cdot W}$ removes the influence of a third variable W from both X and Y. The formula is

$$r_{XY \cdot W} = \frac{r_{XY} - r_{XW}\, r_{WY}}{\sqrt{\left(1 - r_{XW}^2\right)\left(1 - r_{WY}^2\right)}} \qquad \text{(R.15)}$$

Applied to the hypothetical correlations just listed, the partial correlation between foot length and vocabulary breadth controlling for age is $r_{XY \cdot W} = .043$. (You should verify this result.) Because the association between X and Y disappears when W is controlled, their bivariate relation may be spurious. Presumed spurious associations due to common causes are readily represented in SEM.

Equation R.15 for partial correlation can be extended to control for two or more external variables. For example, the **second-order partial correlation** $r_{XY \cdot WZ}$ estimates the association between X and Y controlling for both W and Z. There is a related coefficient called **part correlation** or **semipartial correlation** that controls for external variables out of either of two other variables, but not both. The formula for the **first-order part correlation** $r_{Y(X \cdot W)}$, for which the association between X and W is controlled but not for the association between Y and W, is

$$r_{Y(X\cdot W)} = \frac{r_{XY} - r_{WY}r_{XW}}{\sqrt{1 - r_{XW}^2}} \qquad \text{(R.16)}$$

Given the same bivariate correlations among these three variables reported earlier, the part correlation between vocabulary breadth (Y) and foot length (X) controlling only foot length for age (W) is $r_{Y(X\cdot W)} = .033$. This result (.033) is somewhat smaller than the partial correlation for these data, or $r_{XY\cdot W} = .043$. In general, $r_{XY\cdot W} \geq r_{Y(X\cdot W)}$; if $r_{XW} = 0$, then $r_{XY\cdot W} = r_{Y(X\cdot W)}$.

Relations among the squares of the various correlations just described can be illustrated with a Venn-type diagram like the one in Figure R.3. The circles represent total standardized variances of the criterion Y and predictors X and W. The regions in the figure labeled a–d make up the total standardized variance of Y, so

$$a + b + c + d = 1.0$$

Areas a and b represent the proportions of variance in Y uniquely explained by, respectively, X and W, but area c represents the simultaneous overlap (redundancy) of the predictors with the criterion.[1] Area d represents the proportion of unexplained variance. The squared bivariate correlations of the predictors with the criterion and the overall squared multiple correlation can be expressed as sums of the areas a, b, c, or d in Figure R.3, as follows:

$$r_{XY}^2 = a + c \quad \text{and} \quad r_{WY}^2 = b + c$$

$$R_{Y\cdot X,\,W}^2 = a + b + c = 1.0 - d$$

The squared part correlations match up directly with the unique areas a and b in Figure R.3. Each of these areas also equals the *increase* in the total proportion of explained variance that occurs by adding a second predictor to the equation (i.e., ΔR^2); that is,

$$r_{Y(X\cdot W)}^2 = a = R_{Y\cdot X,\,W}^2 - r_{WY}^2 \qquad \text{(R.17)}$$

$$r_{Y(W\cdot X)}^2 = b = R_{Y\cdot X,\,W}^2 - r_{XY}^2$$

The squared partial correlations correspond to areas a, b, and d in Figure R.3, and each estimates the proportion of variance in the criterion explained by one predictor but not the other. The formulas are

$$r_{XY\cdot W}^2 = \frac{a}{a + d} = \frac{R_{Y\cdot X,\,W}^2 - r_{WY}^2}{1 - r_{WY}^2} \qquad \text{(R.18)}$$

[1] Note that interpretation of the area c in Figure R.3 as a proportion of variance generally holds when all bivariate correlations are positive and there is no suppression. Otherwise, the value c can be a negative, but there is no such thing as a negative proportion of variance.

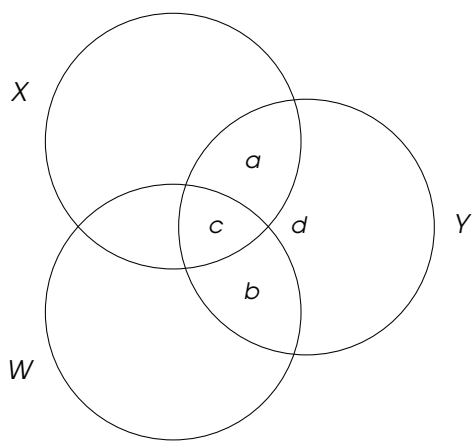


FIGURE R.3. Venn diagram for the standardized variances of predictors *X* and *W* and criterion *Y*.

$$r^2_{WY \cdot X} = \frac{b}{b+d} = \frac{R^2_{Y \cdot X,W} - r^2_{XY}}{1 - r^2_{XY}}$$

For the data in Table R.1, $r^2_{Y(X \cdot W)}$ = .327 and $r^2_{XY \cdot W}$ = .435. In words, predictor X uniquely explains .327, or 32.7% of the total variance of Y (squared part correlation). Of the variance in Y not already explained by W, predictor X accounts for .435, or 43.5% of the remaining variance (squared partial correlation). Exercise 7 asks you to calculate and interpret the corresponding results for the other predictor, W, and the same data.

When predictors are correlated—which is just about always—beta weights, partial correlations, and part correlations are alternative ways to describe in standardized terms the relative explanatory power of each predictor controlling for the rest. None is more "correct" than the others because each gives a different perspective on the same data. Note that unstandardized regression coefficients (B) are preferred when comparing results for the same predictors and criterion across different samples.

OBSERVED VERSUS ESTIMATED CORRELATIONS

The Pearson correlation estimates the degree of linear association between two continuous variables. Its equation is

$$r_{XY} = \frac{\text{cov}_{XY}}{SD_X SD_Y} = \frac{\sum_{i=1}^{N} z_{X_i} z_{Y_i}}{df} \qquad \text{(R.19)}$$

where $df = N - 1$. Rodgers and Nicewander (1988) described a total of 11 other formulas, each of which represents a different conceptual or computational definition of r, but all of which yield the same result for the same data.

A continuous variable is one for which, theoretically, any value is possible within the limits of its score range. This includes values with decimals, such as 3.75 seconds or 13.60 kilograms. In practice, variables with a range of at least 15 points or so are usually considered as continuous even if their scores are discrete, or integers only (e.g., scores of 10, 11, 12, etc.). For example, the PRELIS program of LISREL—used for data preparation—automatically classifies a variable with less than 16 levels as ordinal.

The statistic r has a theoretical maximum absolute value of 1.0. But the practical upper limit for $|r|$ is < 1.0 if the relation between X and Y is not unconditionally linear, there is measurement error in either X or Y, or distributions for X versus Y have different shapes. The amount of variation in samples (i.e., SD_X and SD_Y in Equation R.19) also affects the value of r. In general, restriction of range on either X or Y through sampling or case selection (e.g., only cases with higher scores on X are studied) tends to reduce values of $|r|$, but not always (see Huck, 1992). The presence of outliers, or extreme scores, can also distort the value of r; see Goodwin and Leech (2006) for more information.

There are other forms of the Pearson correlation for observed variables that are either natural dichotomies, such as male versus female for chromosomal sex, or ordinal (ranks). For example:

1. The **point-biserial correlation** (r_{pb}) estimates the association between a dichotomy and a continuous variable (e.g., treatment vs. control, weight).
2. The **phi coefficient** ($\hat{\varphi}$) is for two dichotomies (e.g., treatment vs. control, survived vs. died).
3. **Spearman's rank order correlation** or **Spearman's rho** ($\hat{\rho}$) is for two ranked variables (e.g., finish order in a race, rank by amount of training time).

Computational formulas for all these special forms are just rearrangements of Equation R.19 for r (e.g., Kline, 2013a, pp. 138, 166).

All forms of the Pearson correlation estimate associations between observed (measured) variables. Other, non-Pearson correlations assume that the underlying, or latent, variables are continuous and normally distributed. For example:

1. The **biserial correlation** (r_{bis}) is for a naturally continuous variable, such as weight, and a dichotomy, such as recovered–not recovered, that theoretically represents a dichotomized continuous latent variable. For example, presumably degrees of recovery were collapsed when the observed dichotomy was created. The value of r_{bis} estimates what the Pearson r would be if the dichotomized variable were continuous and normally distributed.
2. The **polyserial correlation** is the generalization of

r_{bis} that does basically the same thing for a naturally continuous variable and a theoretically continuous-but-polytomized variable (i.e., categorized into three or more levels). Likert-type response scales for survey or questionnaire items, such as *agree*, *undecided*, or *disagree*, are examples of a polytomized response continuum about the degree of agreement.

3. The **tetrachoric correlation** (r_{tet}) for two dichotomized variables estimates what r would be if both measured variables were continuous and normally distributed.
4. The **polychoric coefficient** is the generalization of the tetrachoric correlation that estimates r but for ordinal observed variables with two or more levels.

Computing polyserial or polychoric correlations is relatively complicated and requires special software, such as PRELIS in LISREL. These programs generally use a special form of maximum likelihood estimation that assumes normality of the latent continuous variables, and error variance tends to increase rapidly as the number of categories on the observed variables decreases from about five to two; that is, dichotomized continuous variables generate the greatest imprecision.

The PRELIS program can also analyze **censored variables**, for which values occur outside of the range of measurement. Suppose that a scale registers values of weight between 1 and 300 pounds only. For objects that weigh either less than 1 pound or more than 300 pounds, the scale tells us only that the measured weight is, respectively, at most 1 pound or at least 300 pounds. In this example, the hypothetical scale is both left censored and right censored because the values less than 1 or more than 300 are not registered on the scale. There are other possibilities for censoring, but scores on censored variables are either exactly known (e.g., weight = 250) or partially known in that they fall within an interval (e.g., weight ≥ 300). The technique of **censored regression**, better known in economics than in the behavioral sciences, analyzes censored outcomes.

In SEM, Pearson correlations are normally analyzed as part of analyzing covariances when outcome variables are continuous. But noncontinuous outcome variables can be analyzed in SEM, too. One option is to calculate polyserial or polychoric correlations from the raw data and then fit the model to these predicted Pearson correlations. Special methods for analyzing noncontinuous variables in SEM are considered later in Chapters 17 and 18.

In both regression and SEM, it is generally a bad idea to categorize predictors or outcomes that are continuous in order to form **pseudo-groups** (e.g., "low" vs. "high" based on a mean split). Categorization not only discards numerical information about individual differences in the original distribution but it also tends to reduce absolute values of sample correlations when population distributions are normal. The degree of this reduction is greater as the cutting point moves further away from the mean. But if population correlations are low and the sample size is small, then categorization can actually increase absolute sample correlations. Categorization can also create artifactual main or interactive effects, especially when cutting points are arbitrary. In general, it is better to analyze continuous variables as they are and without categorizing them—see Royston, Altman, and Sauerbrei (2006) for more information.

LOGISTIC REGRESSION AND PROBIT REGRESSION

Some options to analyze dichotomous outcomes in SEM are based on **logistic regression**. Just as in standard multiple regression, the predictors in logistic regression can be either continuous or categorical. But the prediction equation in logistic regression is a **logistic function**, or a sigmoid function with an "S" shape. It is a type of **link function**, or a transformation that relates the observed outcomes to the predicted outcomes in a regression analysis. Each method of regression has its own special kind of link function. In standard multiple regression with continuous variables, the link function is the **identity link**, which says that observed scores on the criterion Y are in the same units as $\hat{Y}$, the predicted scores (e.g., Figure R.1). For noncontinuous outcomes, though, original and predicted scores are in different metrics. This is also true in logistic regression, where the link function is the **logit link** as explained next.

Suppose that a total of 32 patients with the same disorder are administered a daily treatment for a varying number of days (5–60). After treatment, the patients are rated as recovered (1) or not recovered (0). Presented in Table R.2 are the hypothetical raw data for this example. I used Statgraphics Centurion (Statgraphics Technologies, 1982–2022)[2] to plot the logistic function with

[2] *https://www.statgraphics.com/centurion-overview*

TABLE R.2. Example Data Set for Logistic Regression and Probit Regression

Status	*n*	Number of days in treatment (*X*)
Not recovered (*Y* = 0)	16	6, 7, 9, 10, 11, 13, 15, 16, 18, 19, 23, 25, 26, 28, 30, 32
Recovered (*Y* = 1)	16	27, 30, 33, 35, 36, 39, 41, 42, 44, 46, 47, 49, 51, 53, 55, 56

95% confidence limits for these data that is presented in Figure R.4. This function generates $\hat{\pi}$, the predicted probability of recovery, given the number of days treated, *X*. The confidence limits for these predictions are so wide because the sample size is small (see the figure). Because predicted probabilities are estimated from the data, they correspond to a latent continuous variable, and in this sense logistic regression (and probit regression, too) can be seen as a latent variable technique.

The estimation method in logistic regression is not OLS. Instead, it is usually a form of maximum likelihood estimation that is applied after transforming the dichotomous outcome variable into a **logit**, which is the natural logarithm (i.e., natural base *e*, or about 2.7183) of the **odds** of the target outcome, $\hat{\omega}$. The quantity $\hat{\omega}$ is the ratio of the probability for the target event, such as recovered, over the probability for the other event, such as not recovered. Suppose that 60% of patients recover after treatment, but the rest, or 40%, do not recover, or

$$\hat{\pi} = .60 \text{ and } 1 - \hat{\pi} = .40$$

The odds of recovery are thus $\hat{\omega} = .60/.40$, or 1.50; that is, the odds are 3:2 in favor of recovery. Odds are converted back to probabilities by dividing the odds by 1.0 plus the odds. For example, $\hat{\omega} = 1.50$, so $\hat{\pi} = 1.50/2.50 = .60$, which is the probability of recovery.

Coefficients for predictors in logistic regression are calculated by the computer in a log metric, but each coefficient can be converted to an **odds ratio**, which

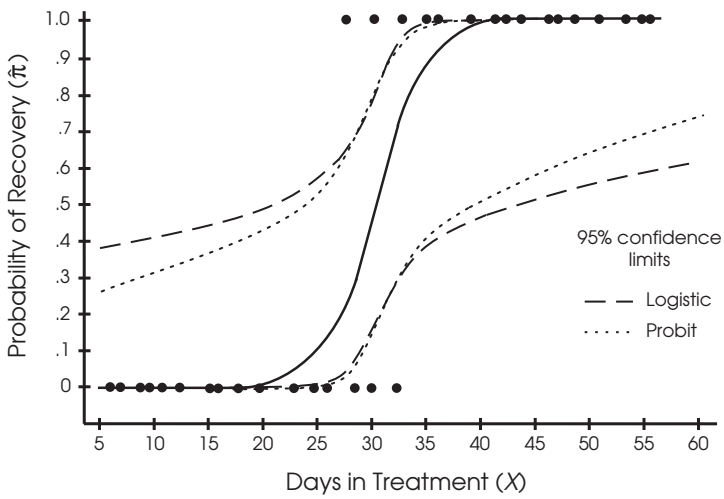


FIGURE R.4. Predicted probability of recovery with 95% confidence limits for the data in Table R.2.

estimates the difference in the odds of the target outcome, given a 1-point increase in the predictor, controlling for all other predictors. I submitted the data in Table R.2 to the Logistic Regression procedure in Statgraphics Centurion. The prediction equation in a log metric is

$$\text{logit}(\hat{\pi}) = \ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = \ln(\hat{\omega}) = .455X - 13.701$$

where .455 is the coefficient for the predictor X, number of treatment days, and –13.701 is the intercept. Taking the antilogarithm of the coefficient for days in treatment, or

$$\ln^{-1}(.455) = e^{.455} = 1.576$$

gives us the odds ratio, or 1.576. This result says that for each additional day of treatment, the odds for recovery increase by 57.6%. But this rate of increase is not linear; instead, the rate at which a logistic curve ascends or descends changes according to values of the predictor. For these data, the greatest rate of change in predicted recovery occurs between 30 and 40 days of treatment. But at the extremes ($X < 30$ or $X > 40$), the rate of change in the probability of recovery is much less—see Figure R.4. The inverse logit function presented next generates the logistic curve plotted in the figure:

$$\hat{\pi} = \text{logit}^{-1}(.455X - 13.701) = \frac{e^{.455\,X - 13.701}}{1 + e^{.455\,X - 13.701}}$$

An alternative method is **probit regression**, which analyzes binary outcomes in terms of a **probit function**, where probit stands for "probability unit." Likewise, the link function in probit regression is the **probit link**. A probit model assumes that the observed dichotomy $Y = 1$ for the target outcome versus $Y = 0$ for other events is determined by a normal continuous latent variable Y^* with a mean of zero and variance of 1.0 such that

$$Y = \begin{cases} 1 & \text{if} \quad Y^* \geq 0 \\ 0 & \text{if} \quad Y^* < 0 \end{cases} \qquad \text{(R.20)}$$

The equation in probit regression generates $\hat{Y}^*$ in the metric of normal deviates (z scores). Next, the computer uses the equation for the cumulative distribution function of the normal curve (Φ) to calculate predicted probabilities of the target outcome $\hat{\pi}$ from values of $\hat{Y}^*$ for each case:

$$\hat{\pi} = \Phi(\hat{Y}^*) \qquad \text{(R.21)}$$

Equation R.21 is known as the **normal ogive model.**[3]

I analyzed the data in Table R.2 using the Probit Analysis procedure in Statgraphics Centurion. The prediction equation is

$$\hat{Y}^* = .268X - 8.072$$

The coefficient for X, .268, estimates in standard deviation units the amount of change in recovery, given a one-day increase in treatment. That is, the z-score for recovery increases by .268 for each additional day of treatment. Again, this rate of change is not constant because the overall relation is nonlinear (Figure R.4). Predicted probabilities of recovery for this example are generated by the probit function

$$\hat{\pi} = \Phi(.268X - 8.072)$$

The 95% confidence limits for the probit function are somewhat different than those for the logistic function for the data in Table R.2—see Figure R.4.

Logistic regression and probit regression applied in the same large samples tend to give similar results but in different metrics for the coefficients. The scaling factor that converts results from the logistic model to the same metric as the normal ogive (probit) model is approximately 1.7. For example, the ratio of the coefficients for the predictor in, respectively, the logistic and probit analyses of the data in Table R.2 is .455/.268 = 1.698, or 1.7 at single-decimal accuracy. The two procedures may generate appreciably different results if there are many cases at the extremes (predicted probabilities are close to either 0 to 1.0) or if the sample is small. Probit regression is more computationally intensive than logistic regression, but this difference is relatively unimportant for modern microcomputers with fast processors and ample memory. It can happen that computer procedures for probit regression may fail to generate a solution in smaller samples. Agresti (2019) describes additional techniques for categorical data.

[3] You can see the equation for Φ at *https://en.wikipedia.org/wiki/Normal_distribution*

SUMMARY

You should know about regression analysis before learning the basics of SEM. For both sets of techniques, the results are affected not only by what is measured (i.e., the data) but also by what is not measured, especially if omitted predictors covary with included predictors, which is a specification error. Accordingly, you should carefully select predictors after review of theory and results of prior studies in the area. In regression, those predictors should have adequate psychometric characteristics because there is no allowance for measurement error. The same restriction does not apply in SEM, but use of grossly inadequate measures in SEM can seriously bias the results, too. When selecting predictors, the role of judgment should be greater than that of significance testing, which can greatly capitalize on sample-specific variation.

LEARN MORE

The book by Cohen, Cohen, West, and Aiken (2003) is considered by many as a kind of "bible" for multiple regression. Royston, Altman, and Sauerbrei (2006) explain why categorizing predictor or outcome variables is a bad idea. Shieh (2006) describes suppression in more detail.

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

Royston, P., Altman, D. G., & Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: A bad idea. *Statistics in Medicine, 25*, 127–141.

Shieh, G. (2006). Suppression situations in multiple linear regression. *Educational and Psychological Measurement, 66*, 435–447.

EXERCISES

All questions concern the data in Table R.1.

1. Calculate the unstandardized regression equation for predicting Y from X based on the descriptive statistics.

2. Show that centering scores on X does not change the value of the unstandardized regression coefficient for predicting Y but does affect the value of the intercept.

3. Show that $s_Y^2 = s_{\hat{Y}}^2 + s_{Y-\hat{Y}}^2$ and $r_{XY}^2 = s_{\hat{Y}}^2 / s_Y^2$ when X is the only predictor of Y.

4. Calculate the unstandardized regression equation and the standardized regression equation for predicting Y from both X and W. Also calculate $R_{Y \cdot X,\, W}^2$.

5. Calculate $\hat{R}_{Y \cdot X,\, W}^2$.

6. Construct a histogram of the residuals for the regression of Y on both X and W.

7. Compute and interpret $r_{WY \cdot X}^2$ and $r_{Y(X \cdot W)}^2$.

ANSWERS

1. Given the descriptive statistics and with slight rounding error:

$$B_X = .686\left(\frac{10.870}{3.007}\right) = 2.479$$

$$A_X = 102.950 - 2.479\ (16.900) = 61.054$$

2. Given $M_X = 16.900$, mean-centered scores (x) are

−.90, −2.90, −.90, −4.90, 1.10,
1.10, −3.90, −.90, 1.10, 5.10,
1.10, 2.10, −.90, −.90, 5.10,
−4.90, 3.10, −2.90, 4.10, .10

and $M_x = 0$, $SD_x = 3.007$, $r_{xY} = .686$, so with slight rounding error

$$B_X = .686\left(\frac{10.870}{3.007}\right) = 2.479$$

$$A_x = 102.950 - 2.479\ (0) = 102.950$$

3. Given $\hat{Y} = 2.479\ X + 61.054$, the predicted scores $\hat{Y}$ are

```
100.719, 95.761, 100.719, 90.803,
105.677, 105.677, 93.282, 100.719,
105.677, 115.593, 105.677, 108.156,
100.719, 100.719, 115.593, 90.803,
110.635, 95.761, 113.114, 103.198
```

and the residual scores $\hat{Y} - Y$ are

```
−.719, −3.761, −12.719, 4.197, −7.677,
−4.677, 3.718, −2.719, 4.323, 8.407,
−3.677, 6.844, −8.719, 1.281, −11.593,
−5.803, 7.365, 9.239, −2.114, 18.802
```

With slight rounding error,

$s^2_Y = s^2_{\hat{Y}} + s^2_{Y-\hat{Y}} = 55.570 + 62.586 = +\ 118.155$

$r^2_{XY} = s^2_Y / s^2_{\hat{Y}} = 55.570/118.155 = .470$, so $r_{XY} = .686$

4. Given the descriptive statistics and with slight rounding error:

$$b_X = \frac{.686 - .499(.272)}{1 - .272^2} = .594$$

and

$$B_X = .594\left(\frac{10.870}{3.007}\right) = 2.147$$

$$b_W = \frac{.499 - .686(.272)}{1 - .272^2} = .337$$

and

$$B_W = .337\left(\frac{10.870}{2.817}\right) = 1.302$$

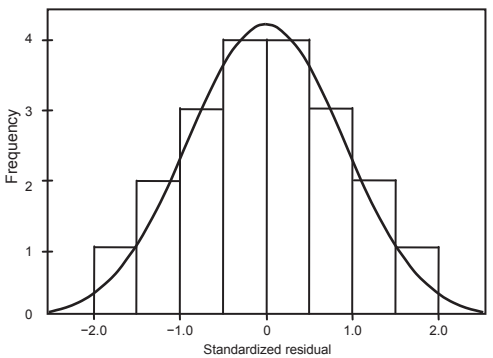
$A_{X,\,W} = 102.950 - 2.147\ (16.900) - 1.302\ (49.400) = 2.340$

$$R^2_{Y \cdot X,W} = .595(.686) + .337(.499) = .576$$

5. For $N = 20$, $k = 2$ and $R^2_{Y \cdot X,W} = .576$:

$$\hat{R}^2_{Y \cdot X,W} = 1 - (1 - .576)\left(\frac{20 - 1}{20 - 2 - 1}\right) = .526$$

6. Presented next is the distribution of standardized residuals for the regression of Y on both X and W generated in SPSS with a superimposed normal curve:



7. For $r_{XY} = .686$, $r_{WY} = .499$, $r_{XW} = .272$, and $R^2_{Y \cdot X,W} = .576$ with slight rounding error:

$$r^2_{Y(W \cdot X)} = .576 - .686^2 = .\frac{(.499 - .686(.272))^2}{1 - .272^2} = .105$$

$$r^2_{WY \cdot X} = \frac{.576 - .686^2}{1 - .686^2} = .\frac{(.499 - .686(.272))^2}{(1 - .272^2)(1 - .686^2)} = .199$$

Respectively, variable W uniquely explains about 10.5% of the total variance in Y, and of variance in Y not already explained by X, predictor W accounts for about 19.9% of the rest.

REFERENCES

Achen, C. H. (2005). Let's put garbage-can regressions and garbage-can probits where they belong. *Conflict Management and Peace Science, 22*(4), 327–339.

Agresti, A. (2019). *An introduction to categorical data analysis* (3rd ed.). Wiley.

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley.

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

Goodwin, L. D., & Leech, N. L. (2006). Understanding correlation: Factors that affect the size of *r*. *Journal of Experimental Education, 74*(3), 251–266.

Huck, S. W. (1992). Group heterogeneity and Pearson's *r*. *Educational and Psychological Measurement, 52*(2), 253–260.

Kenny, D. A., & Milan, S. (2012). Identification: A nontechnical discussion of a technical issue. In R. H. Hoyle, (Ed.), *Handbook of structural equation modeling* (pp. 145–163). Guilford Press.

Kline, R. B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences* (2nd ed.). American Psychological Association.

Loehlin, J. C. (2004). *Latent variable models: An introduction to factor, path, and structural equation analysis* (4th ed.). Erlbaum.

Mauro, R. (1990). Understanding L.O.V.E. (left out variables error): A method for estimating the effects of omitted variables. *Psychological Bulletin, 108*(2), 314–329.

Rodgers, J. L., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *American Statistician, 42*(1), 59–66.

Royston, P., Altman, D.G., & Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: A bad idea. *Statistics in Medicine, 25*(1), 127–141.

Shieh, G. (2006). Suppression situations in multiple linear regression. *Educational and Psychological Measurement, 66*(3), 435–447.

Statgraphics Technologies, Inc. (1982–2013). Statgraphics Centurion (Version 19.4.01). [Computer software]. *https://www.statgraphics.com/*

Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement, 55*(4), 525–534.

Wherry, R. J. (1931). A new formula for predicting the shrinkage of the coefficient of multiple correlation. *Annals of Mathematical Statistics, 2*(4), 440–451.

Whittingham, M. J., Stephens, P. A., Bradbury, R. B., & Freckleton, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology, 75*(5), 1182–1189.

Williams, M. N., Grajales, C. A. G., & Kurkiewicz, D. (2013). Assumptions of multiple regression: Correcting two misconceptions. *Practical Assessment, Research, and Evaluation, 18*, Article 11.