Chapter 7
Interval-Level Variables

Chapter 7 of *The Measurement of Association* applies exact and Monte Carlo permutation statistical methods to measures of association designed for two or more interval-level variables. While permutation statistical methods are commonly associated with non-parametric statistics and, therefore, thought by many to be limited to nominal- and ordinal-level measurements, such is certainly not the case, as noted by Feinstein in 1973 [12]. In fact, a great strength of exact and Monte Carlo permutation statistical methods is in the analysis of interval-level measurements [6]. Chapter 7 begins with a discussion and comparison of simple and multiple ordinary least squares (OLS) regression and simple and multiple least absolute deviation (LAD) regression using permutation statistical methods. Multiple regression with multiple independent variables and multivariate dependent variables is described and illustrated. Point-biserial and biserial correlation coefficients are described and analyzed with exact and Monte Carlo permutation methods. Fisher's z transform is examined and evaluated as to its utility in transforming skewed distributions for both hypothesis testing and confidence intervals. Chapter 7 concludes with a discussion of permutation statistical methods applied to Pearson's intraclass correlation coefficient.

7.1 Ordinary Least Squares (OLS) Linear Regression

Ordinary least squares (OLS) regression with a single predictor is a popular statistical measure of the degree of association (correlation) between two interval-level variables, usually denoted as x and y. The assumption of normality comes into play when the null hypothesis is tested by conventional means. Permutation statistical methods do not assume normality and, therefore, are often more useful than conventional statistical methods, especially when the sample size is small. Let r_{xy} denote the Pearson product-moment correlation coefficient for variables x and y

K. J. Berry et al., *The Measurement of Association*,
https://doi.org/10.1007/978-3-319-98926-6_7

given by

$$r_{xy} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{N}(x_i - \bar{x})^2\right]\left[\sum_{i=1}^{N}(y_i - \bar{y})^2\right]}} ,$$

where $\bar{x}$ and $\bar{y}$ denote the arithmetic means of variables x and y, respectively, and N is the number of bivariate measurements. The conventional test of significance is given by

$$t = \frac{r_{xy}\sqrt{N-2}}{\sqrt{1 - r_{xy}^2}} ,$$

which is distributed as Student's t with $N - 2$ degrees of freedom, under the assumption of normality.

More useful than simple OLS regression and correlation is multiple OLS regression with p predictors, $x_1, x_2, \ldots, x_p$. Let $R_{y.x_1, x_2, \ldots, x_p}$ indicate the multiple correlation coefficient for variables y and $x_1, x_2, \ldots, x_p$ given by

$$R^2_{x_1, x_2, \ldots, x_p} = \boldsymbol{\beta}' \mathbf{r}_y ,$$

where $\boldsymbol{\beta}'$ is the transposed vector of standardized regression weights and $\mathbf{r}_y$ is the vector of zero-order correlation coefficients of y with $x_1, x_2, \ldots, x_p$. The conventional test of significance is given by

$$F = \frac{(N - p - 1)R^2_{y.x_1, x_2, \ldots, x_p}}{p(1 - R^2_{y.x_1, x_2, \ldots, x_p})} ,$$

which is distributed as Snedecor's F with p and $N - p - 1$ degrees of freedom, under the assumption of normality.

7.1.1 Univariate Example of OLS Regression

Consider the example set of bivariate data listed in Table 7.1 for $N = 11$ subjects. For the bivariate data listed in Table 7.1, the Pearson product-moment correlation coefficient is $r_{xy} = +0.8509$. An exact permutation analysis requires random shuffles of either the x or the y values with the other set of values held constant.

Table 7.1 Example bivariate OLS correlation data on $N = 11$ subjects

Subject	x	y
1	11	4
2	18	11
3	12	1
4	27	16
5	15	5
6	21	9
7	25	10
8	15	2
9	18	8
10	23	7
11	12	3

For this small example there are

$$M = N! = 11! = 39{,}916{,}800$$

possible, equally-likely arrangements in the reference set of all permutations of the observed bivariate data, making an exact permutation analysis feasible. Monte Carlo resampling methods are generally preferred for permutation correlation analyses since $N!$ is usually a very large number, e.g., with $N = 13$ there are $13! = 6{,}227{,}020{,}800$ possible arrangements. Let r_o indicate the observed value of r_{xy}. Then, based on $L = 1{,}000{,}000$ random arrangements of the observed data under the null hypothesis, there are 861 $|r_{xy}|$ values equal to or greater than $|r_o| = 0.8509$, yielding a Monte Carlo resampling two-sided probability value of $P = 861/1{,}000{,}000 = 0.8610 \times 10^{-3}$.

While $M = 39{,}916{,}800$ possible arrangements of the observed data makes an exact permutation analysis impractical, it is not impossible. Based on the $M = 39{,}916{,}800$ arrangements of the observed data under the null hypothesis, there are 35,216 $|r_{xy}|$ values equal to or greater than $|r_o| = 0.8509$, yielding an exact two-sided probability value of $P = 35{,}216/39{,}916{,}800 = 0.8822 \times 10^{-3}$. For comparison, for the data listed in Table 7.1 $t = 4.8591$ and the two-sided probability value of $|r_o| = 0.8509$ based on Student's t distribution with $N - 2 = 11 - 2 = 9$ degrees of freedom is $P = 0.8969 \times 10^{-3}$.

7.1.2 Multivariate Example of OLS Regression

For a multivariate example of OLS linear regression, consider the small example data set with $p = 2$ predictors listed in Table 7.2 where variable y is Weight in pounds, variable x_1 is Height in inches, and variable x_2 is Age in years for $N = 12$ school children. For the multivariate data listed in Table 7.2, the unstandardized

Table 7.2 Example multivariate OLS correlation data on $N = 12$ children

Child	x_1	x_2	y
1	57	8	64
2	59	10	71
3	49	6	53
4	62	11	67
5	51	8	55
6	50	7	58
7	55	10	77
8	48	9	57
9	52	6	56
10	42	12	51
11	61	9	76
12	57	9	68

OLS regression coefficients are

$$\hat{\beta}_1 = +1.1973 \quad \text{and} \quad \hat{\beta}_2 = +1.1709 \,,$$

and the squared OLS multiple correlation coefficient is $R^2_{y.x_1,\,x_2} = 0.7301$ (henceforth, simply R^2). An exact permutation analysis of multiple correlation requires random shuffles of either the x or the y values. It is important to note that the predictor variables must be shuffled as a unit, i.e., $x_1, \ldots, x_p$. Otherwise, a researcher may end up with a combination of predictor variables that make no sense, e.g., 4-year-old child, married, with two children. Thus, it is advisable to simply shuffle the y values. Even with this very small example there are

$$M = N! = 12! = 479{,}001{,}600$$

possible, equally-likely arrangements of the observed data, making an exact permutation analysis impractical. Based on $L = 1{,}000{,}000$ random arrangements of the observed data, the Monte Carlo resampling probability of $R^2 = 0.7301$ is

$$P\left(R^2 \geq R^2_{\mathrm{o}} \middle| H_0\right) = \frac{\text{number of } R^2 \text{ values } \geq R^2_{\mathrm{o}}}{L} = \frac{2{,}370}{1{,}000{,}000} = 0.2370 \times 10^{-2} \,,$$

where R^2_{o} denotes the observed value of R^2.

While $M = 479{,}001{,}600$ possible arrangements makes an exact permutation analysis impractical, it is not impossible. If the reference set of all possible permutations of the observed scores in Table 7.2 occur with equal chance, the exact

probability of $R^2 = 0.7301$ under the null hypothesis is

$$P\left(R^2 \geq R_o^2 | H_0\right) = \frac{\text{number of } R^2 \text{ values } \geq R_o^2}{M} = \frac{1{,}147{,}714}{479{,}001{,}600} = 0.2396 \times 10^{-2} ,$$

where R_o^2 denotes the observed value of R^2. For comparison, for the data listed in Fig. 7.2, $F = 12.1728$ and the probability value of $R^2 = 0.7301$ based on Snedecor's F distribution with $p, N - p - 1 = 2, 12 - 2 - 1 = 2, 9$ degrees of freedom is approximately $P = 0.2757 \times 10^{-2}$, under the null hypothesis.

7.2 Least Absolute Deviation (LAD) Regression

Ordinary least squares (OLS) linear regression has long been recognized as a useful tool in many areas of research. The optimal properties of OLS linear regression are well known when the errors are normally distributed. In practice, however, the assumption of normality is rarely justified. Least absolute deviation (LAD) linear regression is often superior to OLS linear regression when the errors are not normally distributed [8, 9, 29, 44, 55]. Estimators of OLS regression parameters can be severely affected by unusual values in either the criterion variable or in one or more of the predictor variables, which is largely due to the weight given to each data point when minimizing the sum of squared errors. In contrast, LAD regression is less sensitive to the effects of unusual values because the errors are not squared. The comparison between OLS and LAD linear regression is analogous to the effect of extreme values on the mean and median as measures of location [8]. In this section, the robust nature of least absolute linear regression is illustrated with a simple example and the effects of distance, leverage, and influence are examined. For clarity and efficiency, the illustration and ensuing discussion are limited to simple linear regression with one predictor variable (x) and one criterion variable (y), with no loss of generality.

Consider N paired x_i and y_i observed values for $i = 1, \ldots, N$. For the OLS regression equation given by

$$\hat{y}_i = \hat{\alpha}_{yx} + \hat{\beta}_{yx} x_i ,$$

where $\hat{y}_i$ is the ith of N predicted criterion values and x_i is the ith of N predictor values, $\hat{\alpha}_{yx}$ and $\hat{\beta}_{yx}$ are the OLS parameter estimates of the intercept (α_{yx}) and slope

(β_{yx}), respectively, and are given by

$$\hat{\beta}_{yx} = \frac{\sum_{i=1}^{N} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \tag{7.1}$$

and

$$\hat{\alpha}_{yx} = \bar{y} - \hat{\beta}_{yx}\bar{x} \ , \tag{7.2}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of variables x and y, respectively. Estimates of OLS regression parameters minimize the sum of the squared differences between the observed and predicted criterion values, i.e.,

$$\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \ .$$

For the LAD regression equation given by

$$\tilde{y}_i = \tilde{\alpha}_{yx} + \tilde{\beta}_{yx} x_i \ ,$$

where $\tilde{y}_i$ is the ith of N predicted criterion values and x_i is the ith of N predictor values, $\tilde{\alpha}_{yx}$ and $\tilde{\beta}_{yx}$ are the LAD parameter estimates of the intercept (α_{yx}) and slope (β_{yx}), respectively.[1] Unlike OLS regression, no simple expressions can be given for $\tilde{\alpha}_{yx}$ and $\tilde{\beta}_{yx}$, as for OLS regression in Eqs. (7.1) and (7.2). However, values for $\tilde{\alpha}_{yx}$ and $\tilde{\beta}_{yx}$ may be found through an efficient linear programming algorithm, such as provided by Barrodale and Roberts [1, 2]. In contrast to estimates of OLS regression parameters, estimates of LAD regression parameters minimize the sum of the absolute differences between the observed and predicted criterion values, i.e.,

$$\sum_{i=1}^{N} |y_i - \tilde{y}_i| \ .$$

[1] In this chapter, a caret (^) over a symbol such as $\hat{\alpha}$ or $\hat{\beta}$ indicates an OLS regression model predicted value of a corresponding population parameter, while a tilde (~) over a symbol such as $\tilde{\alpha}$ or $\tilde{\beta}$ indicates a LAD regression model predicted value of a corresponding population parameter.

It is convenient to have a measure of agreement, not correlation, between the observed and predicted y values. Let

$$\delta = \frac{1}{N} \sum_{i=1}^{N} \left| y_i - \tilde{y}_i \right| .$$

Then, the expected value of δ is given by

$$\mu_\delta = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left| y_i - \tilde{y}_j \right| ,$$

and a measure of agreement between the observed y values and the predicted $\tilde{y}$ values is given by

$$\Re = 1 - \frac{\delta}{\mu_\delta} .$$

$\Re$ is a chance-corrected measure of agreement and/or effect size, reflecting the amount of agreement in excess of what would be expected by chance. $\Re$ attains a maximum value of unity when the agreement between the observed y values and the predicted $\tilde{y}$ values is perfect, i.e., y_i and $\tilde{y}_i$ values are identical for $i = 1, \ldots, N$. $\Re$ is zero when the agreement between the observed y values and predicted $\tilde{y}$ values is equal to what is expected by chance, i.e., $\mathrm{E}[\Re | H_0] = 0$. Like all chance-corrected measures, $\Re$ will occasionally be slightly negative when agreement is less than what is expected by chance.

7.2.1 Illustration of Effects of Extreme Values

Three useful diagnostics for assessing the potential effects of extreme values on regression estimators are distance, leverage, and influence. In general terms, *distance* refers to the possible presence of unusual values in the criterion variable and is typically measured as the deviation of a value from the measured center of the criterion variable (y). *Leverage* refers to the possible presence of unusual values in a predictor variable. In the case of a single predictor, leverage is typically measured as the deviation of a value from the measured center of the predictor variable (x). *Influence* incorporates both distance and leverage and refers to the possible presence of unusual values in some combination of the criterion and predictor variables.

For OLS regression, the measure of distance for any data point is simply an error term or residual, i.e., $e_i = y_i - \hat{y}_i$ and is sometimes standardized and sometimes Studentized. Leverage is a measure of the importance of the ith observation in determining the model fit and is usually designated as h_i. More specifically, h_i is

the ith diagonal element of the $N \times N$ matrix

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$$

called the "hat matrix," since $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ in which $\hat{\mathbf{y}}$ is the transposed column vector

$$\hat{\mathbf{y}} = \left(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N\right)' \quad \text{and} \quad \mathbf{y} = (y_1, y_2, \ldots, y_N)' \ .$$

In the case of only one predictor, leverage is simply a function of the deviation of an x score on that predictor from the prediction mean and is given by

$$h_i = \frac{1}{N} + \frac{(x_i - \bar{x})^2}{(N-1)s_x^2} \qquad \text{for } i = 1, \ldots, N \ ,$$

where s_x^2 is the estimated population variance for variable x given by

$$s_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2 \ .$$

Influence combines both leverage and distance, measured as a Studentized residual, to identify unusually influential observations. Residuals are sometimes standardized and sometimes Studentized. Standardized residuals are given by

$$z_i = \frac{e_i}{s_{y.x}} \qquad \text{for } i = 1, \ldots, N \ ,$$

where $e_i = y_i - \hat{y}_i$ for $i = 1, \ldots, N$ is the unstandardized residual and

$$s_{y.x} = \left(\frac{1}{N-p-1}\sum_{i=1}^{N} e_i^2\right)^{1/2}$$

is the standard error of estimate. Standardized residuals have a mean of zero and a variance of one. Studentized residuals are given by

$$r_i = \frac{e_i}{s_{y.x}\sqrt{1-h_i}} = \frac{z_i}{\sqrt{1-h_i}} \qquad \text{for } i = 1, \ldots, N \ .$$

Studentized residuals follow Student's t distribution with mean near zero and variance slightly greater than one.

The most common measure of influence is Cook's distance given by

$$d_i = \left(\frac{1}{p+1}\right) r_i^2 \left(\frac{h_i}{1-h_i}\right) \ ,$$

where r_i^2 denotes the squared Studentized residual and p is the number of predictor variables.

To illustrate the effects of extreme values on the estimates of OLS and LAD regression parameters, consider an example of linear regression with one predictor and a single extreme data point. This simplified example permits the isolation and assessment of distance, leverage, and influence and allows comparison of the effects of an atypical value on estimates of OLS and LAD regression parameters. The data for a linear regression with one predictor variable are listed in Table 7.3. The bivariate data listed in Table 7.3 consist of nine data points with $x_i = i$ and $y_i = 10 - i$ for $i = 1, \ldots, 9$ and describe a perfect negative linear relationship. Figure 7.1 displays the example bivariate data listed in Table 7.3 and indicates the directions of unusual values implicit in distance (D), leverage (L), and influence (I).

Table 7.3 Example bivariate data on $N = 9$ objects for a perfect negative linear regression with one predictor variable

	Object								
Variable	1	2	3	4	5	6	7	8	9
x	3	6	1	8	5	9	2	4	7
y	7	4	9	2	5	1	8	6	3

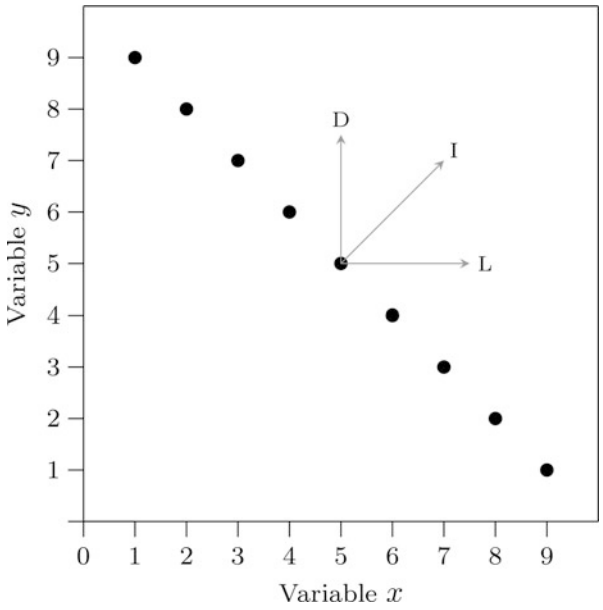

Fig. 7.1 Scatterplot of the data given in Table 7.3 with the directions of extreme values indicated by D, I, and L for distance, influence, and leverage, respectively

Distance

If a tenth bivariate value is added to the nine bivariate values given in Table 7.3 where $(x_{10}, y_{10}) = (5, 5)$, the new data point is located at the common mean and median of both variable x and variable y and, therefore, does not affect the perfect linear relationship between the variables. If x_{10} is held constant at $x_{10} = 5$, but y_{10} takes on the added values of 6, 7, ..., 30, 40, 60, 80, and 100, then the effects of distance on the two regression models can be observed. The vertical movement of y_{10} with variable x held constant at $x_{10} = 5$ is depicted by the directional arrow labeled "D" in Fig. 7.1 and by the four white circles in Fig. 7.2, illustrating an additional data point moving vertically away from location $(x_5, y_5) = (5, 5)$ by increments of one y unit, i.e., (5, 6), (5, 7), (5, 8), and so on.

Table 7.4 lists the values for x_{10} and y_{10} in the first two columns, the $\hat{\alpha}_{yx}$ and $\hat{\beta}_{yx}$ estimates of the OLS regression parameters in the next two columns, and the $\tilde{\alpha}_{yx}$ and $\tilde{\beta}_{yx}$ estimates of the LAD regression parameters in the last two columns. The $\tilde{\alpha}_{yx}$ and $\tilde{\beta}_{yx}$ parameter estimates in the last two columns of Table 7.4 were obtained using the linear program of Barrodale and Roberts [2]. The estimates of the OLS regression parameters listed in Table 7.4 demonstrate that $\hat{\alpha}_{yx}$ systematically changes with increases in distance, but $\hat{\beta}_{yx}$ remains constant at -1.00. In contrast, estimates of the LAD regression parameters are unaffected by changes in distance, remaining constant at $\tilde{\alpha}_{yx} = 10.00$ and $\tilde{\beta}_{yx} = -1.00$ for $x_{10} = 5$ and any value of y_{10}. Given the nine bivariate data points listed in Table 7.3 and an additional

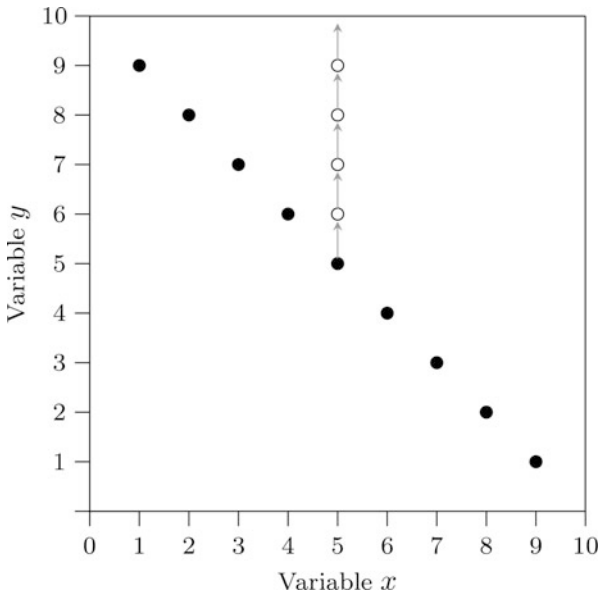

Fig. 7.2 Scatterplot of the data given in Table 7.3 with the locations of an added tenth value indicated by four white circles

Table 7.4 Effects of distance on intercepts and slopes of OLS and LAD linear regression models

		OLS model		LAD model	
x_{10}	y_{10}	$\hat{\alpha}_{yx}$	$\hat{\beta}_{yx}$	$\tilde{\alpha}_{yx}$	$\tilde{\beta}_{yx}$
5	5	+10.0000	−1.0000	+10.0000	−1.0000
5	6	+10.1000	−1.0000	+10.0000	−1.0000
5	7	+10.2000	−1.0000	+10.0000	−1.0000
5	8	+10.3000	−1.0000	+10.0000	−1.0000
5	9	+10.4000	−1.0000	+10.0000	−1.0000
5	10	+10.5000	−1.0000	+10.0000	−1.0000
5	11	+10.6000	−1.0000	+10.0000	−1.0000
5	12	+10.7000	−1.0000	+10.0000	−1.0000
5	13	+10.8000	−1.0000	+10.0000	−1.0000
5	14	+10.9000	−1.0000	+10.0000	−1.0000
5	15	+11.0000	−1.0000	+10.0000	−1.0000
5	16	+11.1000	−1.0000	+10.0000	−1.0000
5	17	+11.2000	−1.0000	+10.0000	−1.0000
5	18	+11.3000	−1.0000	+10.0000	−1.0000
5	19	+11.4000	−1.0000	+10.0000	−1.0000
5	20	+11.5000	−1.0000	+10.0000	−1.0000
5	21	+11.6000	−1.0000	+10.0000	−1.0000
5	22	+11.7000	−1.0000	+10.0000	−1.0000
5	23	+11.8000	−1.0000	+10.0000	−1.0000
5	24	+11.9000	−1.0000	+10.0000	−1.0000
5	25	+12.0000	−1.0000	+10.0000	−1.0000
5	26	+12.1000	−1.0000	+10.0000	−1.0000
5	27	+12.2000	−1.0000	+10.0000	−1.0000
5	28	+12.3000	−1.0000	+10.0000	−1.0000
5	29	+12.4000	−1.0000	+10.0000	−1.0000
5	30	+12.5000	−1.0000	+10.0000	−1.0000
5	40	+13.5000	−1.0000	+10.0000	−1.0000
5	60	+15.5000	−1.0000	+10.0000	−1.0000
5	80	+17.5000	−1.0000	+10.0000	−1.0000
5	100	+19.5000	−1.0000	+10.0000	−1.0000

bivariate data point with $x_{10} = 5$, it follows that

$$\sum_{i=1}^{10} |y_i - \tilde{y}_i| = |y_{10} - 5| \, .$$

Leverage

If a tenth bivariate value is added to the nine bivariate values given in Table 7.3 where $y_{10} = 5$ and x_{10} takes on the added values of 6, 7, ..., 30, 40, 60, 80, and 100, then the effects of leverage on the two regression models can be observed.

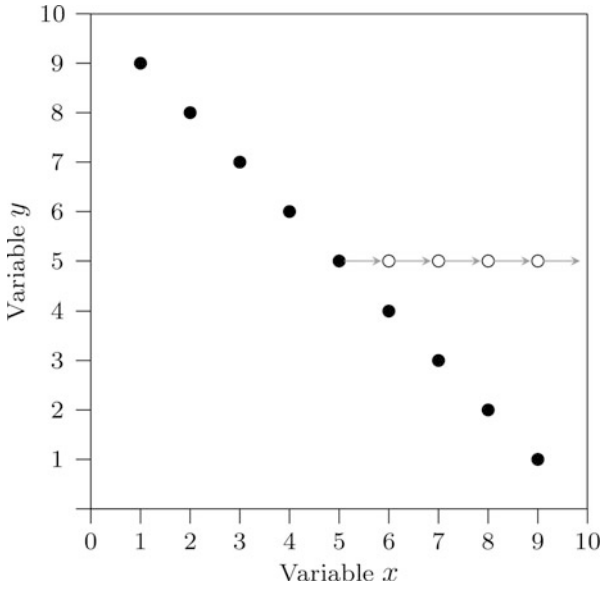

Fig. 7.3 Scatterplot of the data given in Table 7.3 with the locations of an added tenth value indicated by four white circles

The horizontal movement of x_{10} with y_{10} held constant at $y_{10} = 5$ is depicted by the directional arrow labeled "L" in Fig. 7.1 and by the four white circles in Fig. 7.3, illustrating an additional data point moving horizontally away from $(x_5, y_5) = (5, 5)$ by increments of one x unit, i.e., (6, 5), (7, 5), (8, 5), and so on.

Table 7.5 lists the values of x_{10} and y_{10} in the first two columns, the $\hat{\alpha}_{yx}$ and $\hat{\beta}_{yx}$ estimates of the OLS regression parameters in the next two columns, and the $\tilde{\alpha}_{yx}$ and $\tilde{\beta}_{yx}$ estimates of the LAD regression parameters in the last two columns. The $\tilde{\alpha}_{yx}$ and $\tilde{\beta}_{yx}$ estimates were again obtained using the linear program of Barrodale and Roberts [2]. The estimates of the OLS regression parameters listed in Table 7.5 demonstrate that both $\hat{\alpha}_{yx}$ and $\hat{\beta}_{yx}$ exhibit complex changes with increases in leverage. Note the dramatic changes in the intercept from $\hat{\alpha}_{yx} = +10.00$ to $\hat{\alpha}_{yx} = +5.1063$, approaching the mean of y (+5.00), and the slope from $\hat{\beta}_{yx} = -1.00$ to $\hat{\beta}_{yx} = -0.0073$, approaching a slope of zero. In contrast, $\tilde{\alpha}_{yx}$ and $\tilde{\beta}_{yx}$ are unaffected for $y_{10} = 5$ and $5 \leq x_{10} \leq 24$. For $y_{10} = 5$ and $x_{10} \geq 26$, the LAD estimated regression parameters change from $\tilde{\alpha}_{yx} = +10.00$ and $\tilde{\beta}_{yx} = -1.00$ to $\tilde{\alpha}_{yx} = +5.00$ and $\tilde{\beta}_{yx} = 0.00$.

Given the bivariate data listed in Table 7.3 on p. 379 and an additional bivariate data point with variable y held constant at $y_{10} = 5$, it follows that

$$\sum_{i=1}^{10} |y_i - \tilde{y}_i| \leq 20.00$$

Table 7.5 Effects of leverage on intercepts and slopes of OLS and LAD linear regression models

		OLS model		LAD model	
x_{10}	y_{10}	$\hat{\alpha}_{yx}$	$\hat{\beta}_{yx}$	$\tilde{\alpha}_{yx}$	$\tilde{\beta}_{yx}$
5	5	+10.0000	−1.0000	+10.0000	−1.0000
6	5	+10.0246	−0.9852	+10.0000	−1.0000
7	5	+9.9057	−0.9434	+10.0000	−1.0000
8	5	+9.6696	−0.8811	+10.0000	−1.0000
9	5	+9.3548	−0.8065	+10.0000	−1.0000
10	5	+9.0000	−0.7273	+10.0000	−1.0000
11	5	+8.6364	−0.6494	+10.0000	−1.0000
12	5	+8.2853	−0.5764	+10.0000	−1.0000
13	5	+7.9592	−0.5102	+10.0000	−1.0000
14	5	+7.6637	−0.4515	+10.0000	−1.0000
15	5	+7.4000	−0.4000	+10.0000	−1.0000
16	5	+7.1670	−0.3552	+10.0000	−1.0000
17	5	+6.9620	−0.3165	+10.0000	−1.0000
18	5	+6.7822	−0.2829	+10.0000	−1.0000
19	5	+6.6244	−0.2538	+10.0000	−1.0000
20	5	+6.4857	−0.2286	+10.0000	−1.0000
21	5	+6.3636	−0.2066	+10.0000	−1.0000
22	5	+6.2559	−0.1874	+10.0000	−1.0000
23	5	+6.1604	−0.1706	+10.0000	−1.0000
24	5	+6.0756	−0.1559	+10.0000	−1.0000
25	5	+6.0000	−0.1429	+10.0000	−1.0000
26	5	+5.9324	−0.1313	+5.0000	0.0000
27	5	+5.8717	−0.1211	+5.0000	0.0000
28	5	+5.8170	−0.1119	+5.0000	0.0000
29	5	+5.7676	−0.1037	+5.0000	0.0000
30	5	+5.7229	−0.0964	+5.0000	0.0000
40	5	+5.4387	−0.0516	+5.0000	0.0000
60	5	+5.2264	−0.0216	+5.0000	0.0000
80	5	+5.1464	−0.0117	+5.0000	0.0000
100	5	+5.1063	−0.0073	+5.0000	0.0000

for $x_{10} \leq 25$ and

$$\sum_{i=1}^{10} |y_i - \tilde{y}_i| = 20.00$$

for $x_{10} \geq 25$. When $x_{10} \leq 25$, the LAD regression line defined by $\tilde{\alpha}_{yx} = +10.00$ and $\tilde{\beta}_{yx} = -1.00$ yields the minimum sum of absolute differences. However, when $x_{10} \geq 25$ the LAD regression line defined by $\tilde{\alpha}_{yx} = +5.00$ and $\tilde{\beta}_{yx} = 0.00$ that passes through the data point located at (x_{10}, y_{10}) yields the minimum sum of absolute differences. For $x_{10} = 25$, the LAD regression line is not unique. While

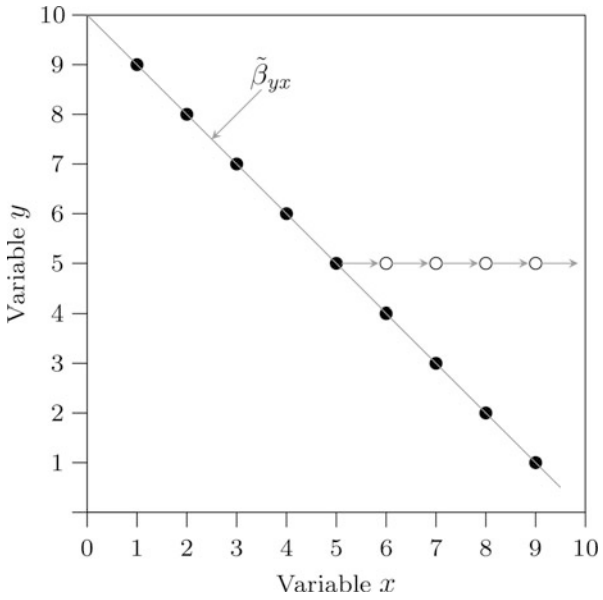

Fig. 7.4 Scatterplot of the data given in Table 7.3 with the regression line $\tilde{\beta}_{yx}$ depicted and the locations of an added tenth value indicated by four white circles

this is an interesting property of LAD regression and can easily be demonstrated with one predictor and a small number of data points, in practice any extreme value would have to be so far removed from the measured center of the distribution of variable x to be considered a "grossly aberrant" value [47, p. 871].

The fact that when $y_{10} = 5$ and $x_{10} = 25$, the solution is not unique and either of the two LAD regression lines is appropriate, deserves some additional explanation. Consider the data points in Fig. 7.4 where the additional tenth point is indicated at locations

$$(x_6, y_5), (x_7, y_5), \ldots, (x_9, y_5)$$

and the LAD regression line for the original nine data points with $\tilde{\alpha} = +10.00$ and $\tilde{\beta} = -1.00$ is depicted. If only the original nine data points are considered, the sum of absolute deviations is zero, i.e.,

$$\sum_{i=1}^{9} |y_i - \tilde{y}_i| = |9-9| + |8-8| + |7-7| + |6-6| + |5-5| + |4-4| + |3-3| + |2-2| + |1-1| = 0.00\,.$$

The addition of a tenth data point at location (x_6, y_5), the first white circle to the right of the regression line in Fig. 7.4, increases the sum of absolute deviations by one, i.e., $|y_i - \hat{y}_i| = |6-5| = 1$. Moving the new data point horizontally to location (x_7, y_5), the second white circle to the right of the regression line in

Fig. 7.4, increases the sum of absolute deviations by two, i.e., $|y_i - \tilde{y}_i| = |7 - 5| = 2$. Continuing to move the new data point horizontally increments the sum of absolute deviations by increasing amounts. Consider locations (x_{24}, y_5), (x_{25}, y_5), and (x_{26}, y_5), where

$$\left|y_i - \tilde{y}_i\right| = \left|24 - 5\right| = 19 \,, \quad \left|y_i - \tilde{y}_i\right| = \left|25 - 5\right| = 20 \,,$$

and

$$\left|y_i - \tilde{y}_i\right| = \left|26 - 5\right| = 21 \,,$$

respectively.

Thus, for an additional value up to location (x_{25}, y_5) the sum of absolute deviations will be equal to or less than 20, and for an additional value beyond location (x_{25}, y_5) the sum of absolute deviations will be equal to or greater than 20. However, when a data point is added at location (x_{25}, y_5) something interesting happens, which is readily apparent in Table 7.5. At this point a dramatic shift in the LAD regression line occurs, from $\tilde{\alpha}_{yx} = +10.00$ and $\tilde{\beta}_{yx} = -1.00$ to $\tilde{\alpha}_{yx} = +5.00$ and $\tilde{\beta}_{yx} = 0.00$. The regression line is leveraged and forced through the new data point location at (x_{25}, y_5). The new regression line is depicted in Fig. 7.5 with the absolute errors indicated by dashed lines. The sum of the absolute errors around the

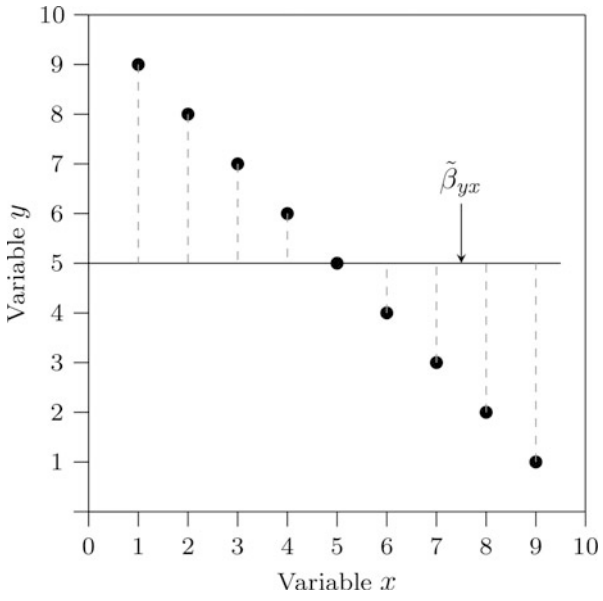

Fig. 7.5 Scatterplot of the data given in Table 7.3 with absolute errors indicated by dashed lines

new regression line is

$$\sum_{i=1}^{10} |y_i - \tilde{y}_i| = |9-5| + |8-5| + |7-5| + |6-5| + |5-5| + |4-5| + |3-5| + |2-5| + |1-5| + |5-5| = 20.00\,.$$

Thus both regression lines given by $\tilde{\alpha}_{yx} = +10.00$ and $\tilde{\beta}_{yx} = -1.00$ and $\tilde{\alpha}_{yx} = +5.00$ and $\tilde{\beta}_{yx} = 0.00$ minimize the sum of absolute deviations when an additional data point is located at (x_{25}, y_5). Note, however, that the additional data point is far to the right and is a very extreme value, unlikely to be encountered in everyday research. Specifically, for this minimalist example, a tenth value at location (x_{25}, y_5) is almost three times the range and over seven standard deviations above the mean—too extreme to be of concern in practice. Thus, LAD regression is highly stable under all but the most extreme cases.

Influence

If a tenth bivariate value is added to the nine bivariate values given in Table 7.3 on p. 379 where $x_{10} = y_{10}$ takes on the added values of 6, 7, …, 30, 40, 60, 80, and 100, then the effects of influence on the two regression models can be observed. The diagonal movement of (x_{10}, y_{10}) is depicted by the directional arrow labeled "I" in Fig. 7.3 and by the four white circles in Fig. 7.6, illustrating an additional data point

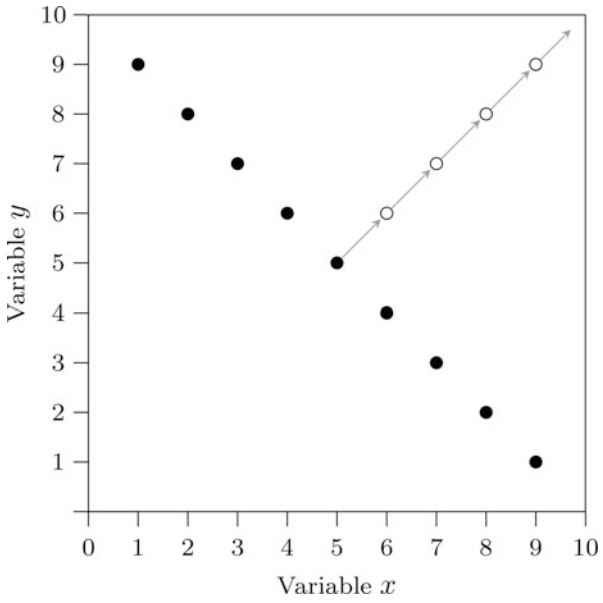

Fig. 7.6 Scatterplot of the data given in Table 7.3 with the locations of an added tenth value indicated by four white circles

moving diagonally away from $(x_5, y_5) = (5, 5)$ by increments of one x and one y unit, i.e., (6, 6), (7, 7), (8, 8), and so on.

Table 7.6 lists the values of x_{10} and y_{10} in the first two columns, the $\hat{\alpha}_{yx}$ and $\hat{\beta}_{yx}$ estimates of the OLS regression parameters in the next two columns, and the $\tilde{\alpha}_{yx}$ and $\tilde{\beta}_{yx}$ estimates of the LAD regression parameters in the last two columns. The estimates of the OLS regression parameters listed in Table 7.4 demonstrate that both $\hat{\alpha}_{yx}$ and $\hat{\beta}_{yx}$ exhibit complex changes with increases in influence, quickly becoming unstable with changes in the intercept from $\hat{\alpha}_{yx} = +10.00$ to $\hat{\alpha}_{yx} = +0.2126$ and changes in the slope from $\hat{\beta}_{yx} = -1.00$ to $\hat{\beta}_{yx} = +0.9853$. Note that $\hat{\beta}_{yx}$ is negative from $x_{10} = 5$ up to $x_{10} = 13$, then changes to positive for $x_{10} = 14$ up to $x_{10} = 100$.

Table 7.6 Effects of influence on intercepts and slopes of OLS and LAD linear regression models

		OLS model		LAD model	
x_{10}	y_{10}	$\hat{\alpha}_{yx}$	$\hat{\beta}_{yx}$	$\tilde{\alpha}_{yx}$	$\tilde{\beta}_{yx}$
5	5	+10.0000	−1.0000	+10.0000	−1.0000
6	6	+10.0493	−0.9704	+10.0000	−1.0000
7	7	+9.8113	−0.8868	+10.0000	−1.0000
8	8	+9.3392	−0.7621	+10.0000	−1.0000
9	9	+8.7097	−0.6129	+10.0000	−1.0000
10	10	+8.0000	−0.4545	+10.0000	−1.0000
11	11	+7.2727	−0.2987	+10.0000	−1.0000
12	12	+6.5706	−0.1527	+10.0000	−1.0000
13	13	+5.9184	−0.0204	+10.0000	−1.0000
14	14	+5.3273	+0.0971	+10.0000	−1.0000
15	15	+4.8000	+0.2000	+10.0000	−1.0000
16	16	+4.3339	+0.2895	+10.0000	−1.0000
17	17	+3.9241	+0.3671	+10.0000	−1.0000
18	18	+3.5644	+0.4342	+10.0000	−1.0000
19	19	+3.2487	+0.4924	+10.0000	−1.0000
20	20	+2.9714	+0.5429	+10.0000	−1.0000
21	21	+2.7273	+0.5868	+10.0000	−1.0000
22	22	+2.5117	+0.6251	+10.0000	−1.0000
23	23	+2.3208	+0.6587	+10.0000	−1.0000
24	24	+2.1512	+0.6882	+10.0000	−1.0000
25	25	+2.0000	+0.7143	0.0000	+1.0000
26	26	+1.8647	+0.7374	0.0000	+1.0000
27	27	+1.7433	+0.7579	0.0000	+1.0000
28	28	+1.6340	+0.7762	0.0000	+1.0000
29	29	+1.5353	+0.7925	0.0000	+1.0000
30	30	+1.4458	+0.8072	0.0000	+1.0000
40	40	+0.8774	+0.8968	0.0000	+1.0000
60	60	+0.4528	+0.9569	0.0000	+1.0000
80	80	+0.2928	+0.9766	0.0000	+1.0000
100	100	+0.2126	+0.9853	0.0000	+1.0000

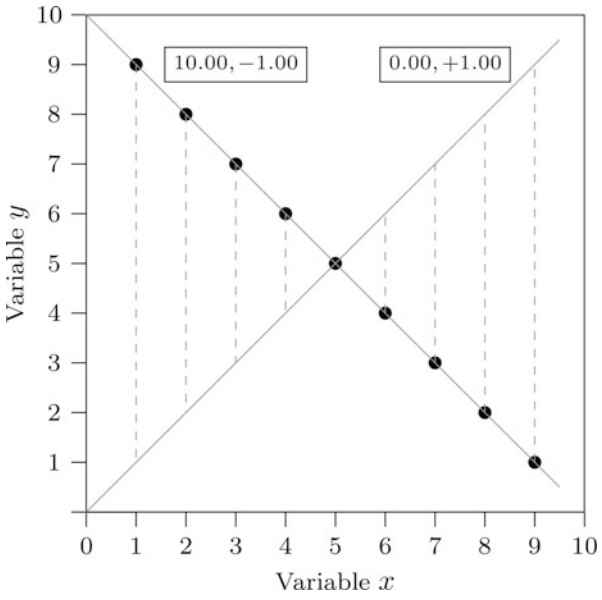

Fig. 7.7 Scatterplot of the data given in Table 7.3 with the regression lines minimizing the sum of absolute errors

Note also that the range of changes in $\hat{\beta}_{yx}$ is from $\hat{\beta}_{yx} = -1.00$ for $x_{10} = 5$ approaching $\hat{\beta}_{yx} = +1.00$ for $x_{10} = 100$; actually, $\hat{\beta}_{yx} = +0.9853$ for $x_{10} = 100$. In contrast, $\tilde{\alpha}_{yx}$ and $\tilde{\beta}_{yx}$ do not change for $5 \leq x_{10} = y_{10} \leq 24$. For $x_{10} = y_{10} \geq 26$, the estimates of the LAD regression parameters change from $\tilde{\alpha}_{yx} = +10.00$ and $\tilde{\beta}_{yx} = -1.00$ to $\tilde{\alpha}_{yx} = 0.00$ and $\tilde{\beta}_{yx} = +1.00$. When $x_{10} = y_{10} = 25$, either of the two LAD regression lines holds since the solution is not unique. Thus, two LAD regression lines minimize the sum of absolute errors: one with $\tilde{\alpha}_{yx} = +10.00$ and $\tilde{\beta}_{yx} = -1.00$ and the other with $\tilde{\alpha}_{yx} = 0.00$ and $\tilde{\beta}_{yx} = +1.00$.

Figure 7.7 depicts the two LAD regression lines, labeled with the values for $\tilde{\alpha}_{yx}$ and $\tilde{\beta}_{yx}$, and dashed lines indicating the errors around the regression line with $\tilde{\alpha}_{yx} = 0.00$ and $\tilde{\beta}_{yx} = +1.00$. As shown in Fig. 7.7, the sum of absolute errors is

$$\sum_{i=1}^{10} |y_i - \tilde{y}_i| = |9-1| + |8-2| + |7-3| + |6-4| + |5-5| + |4-6| + |3-7| + |2-8| + |1-9| + |25-25| = 40.00 .$$

Given the bivariate data listed in Table 7.3 on p. 379 and an additional bivariate data point $x_{10} = y_{10}$, it follows that

$$\sum_{i=1}^{10} |y_i - \tilde{y}_i| \leq 40.00$$

for $5 \leq x_{10} = y_{10} \leq 25$ and

$$\sum_{i=1}^{10} \left| y_i - \tilde{y}_i \right| = 40.00$$

for $x_{10} = y_{10} \geq 25$. When $x_{10} = y_{10} \leq 25$, the LAD regression line defined by $\tilde{\alpha}_{yx} = +10.00$ and $\tilde{\beta}_{yx} = -1.00$ yields the minimum sum of absolute differences between y_i and $\tilde{y}_i$ for $i = 1, \ldots, N$. However, when $x_{10} = y_{10} \geq 25$, the LAD regression line defined by $\tilde{\alpha}_{yx} = 0.00$ and $\tilde{\beta}_{yx} = +1.00$ that passes through the data point located at (x_{10}, y_{10}) yields the minimum sum of absolute differences between y_i and $\tilde{y}_i$ for $i = 1, \ldots, N$. For $x_{10} = y_{10} = 25$, the LAD regression line is not unique. It should be noted that the shift in the LAD regression line is a consequence of only the leverage component of influence. For these data, the LAD regression line is defined by $\tilde{\alpha}_{yx} = +10.00$ and $\tilde{\beta}_{yx} = -1.00$ if $|x_{10} - 5| \leq 20.00$ and the regression line is unique if $|x_{10} - 5| < 20.0$ or $y_{10} = 10 - x_{10}$.

LAD linear regression is a robust alternative to OLS linear regression, especially when errors are generated by fat-tailed distributions [10, 52]. Fat-tailed distributions mean an abundance of extreme values and OLS linear regression gives disproportionate weight to extreme values. In practice, LAD linear regression is virtually unaffected by the presence of a few extreme values. While the effects of distance, leverage, and influence are illustrated with only a simplified example of perfect linear regression with one predictor, the results extend to more general regression models. If a less-than-perfect regression model with p predictors is considered, then the estimators of the LAD regression parameters are unaffected by unusual y_i values, when the leverage effect is absent. In addition, only exceedingly extreme values of the predictors $x_1, \ldots, x_p$ have any effect on the estimation of the LAD regression parameters.

7.2.2 Univariate Example of LAD Regression

Consider the small example set of bivariate data listed in Table 7.7 for $N = 10$ subjects. For the bivariate data listed in Table 7.7, the LAD regression coefficient is $\tilde{\beta} = +2.1111$, $\delta = 5.9889$, $\mu_\delta = 9.2267$, and the LAD chance-corrected measure of agreement between the observed y values and the predicted $\tilde{y}$ values is

$$\Re = 1 - \frac{\delta}{\mu_\delta} = 1 - \frac{5.9889}{9.2267} = +0.3509 \, .$$

Since there are $M = N! = 10! = 3{,}628{,}800$ possible arrangements of the observed data, an exact permutation analysis may not be practical. Based on $L = 1{,}000{,}000$ random arrangements of the observed data, the Monte Carlo resampling probability

Table 7.7 Example bivariate LAD correlation data on $N = 10$ subjects

Subject	x	y
1	14	25
2	8	23
3	5	21
4	2	10
5	1	12
6	3	11
7	9	19
8	2	13
9	3	13
10	9	16

value of $\Re = +0.3509$ is

$$P\left(\Re \geq \Re_{\mathrm{o}} | H_0\right) = \frac{\text{number of } \Re \text{ values } \geq \Re_{\mathrm{o}}}{L} = \frac{6{,}679}{1{,}000{,}000} = 0.6679 \times 10^{-2},$$

where $\Re_{\mathrm{o}}$ denotes the observed value of $\Re$.

While $M = 3{,}628{,}800$ possible arrangements makes an exact permutation analysis impractical, it is not impossible. If the reference set of all possible permutations of the observed scores in Table 7.7 occur with equal chance, the exact probability of $\Re = +0.3509$ under the null hypothesis is

$$P\left(\Re \geq \Re_{\mathrm{o}} | H_0\right) = \frac{\text{number of } \Re \text{ values } \geq \Re_{\mathrm{o}}}{M} = \frac{26{,}966}{3{,}628{,}800} = 0.7431 \times 10^{-2},$$

where $\Re_{\mathrm{o}}$ denotes the observed value of $\Re$.

7.2.3 Multivariate Example of LAD Regression

To illustrate a multivariate LAD linear regression analysis, an application of the LAD regression model to forecasting African rainfall in the western Sahel is utilized [38]. For the multivariate data listed in Table 7.8, the first column lists $N = 15$ calendar years from 1950 to 1964 and the second through fourth columns (U_{50}, U_{30}, and $|U_{50} - U_{30}|$) contain values based on the quasibiennial oscillation of equatorial east/west winds. U_{50} is the zonal wind measured in meters per second at 50 millibars (approximately 20 km in altitude) and U_{30} is the zonal wind measured

Table 7.8 Regional rainfall precipitation by years with predictors U_{50}, U_{30}, $|U_{50} - U_{30}|$, R_s, and R_g

	Predictor					
Year	U_{50}	U_{30}	$\lvert U_{50} - U_{30}\rvert$	R_s	R_g	Rainfall
1950	−3	−3	0	−0.14	+1.07	+1.05
1951	−4	−13	9	+1.68	−0.66	+0.74
1952	−23	−26	3	+0.49	+0.65	+1.45
1953	0	−18	18	+0.93	+0.41	+0.99
1954	−23	−32	9	+0.20	−0.16	+1.12
1955	0	−4	4	+0.60	+0.64	+1.07
1956	−19	−33	14	+1.00	+0.41	+0.36
1957	−2	−3	1	+0.47	−0.36	+0.87
1958	−12	−28	16	+0.58	+1.03	+0.86
1959	−9	−5	4	+1.45	−0.74	+0.30
1960	−6	−21	15	+0.25	+0.12	+0.24
1961	−3	−3	0	+0.23	+1.05	+0.20
1962	−12	−32	20	+0.48	−0.74	+0.41
1963	−17	−3	14	+0.28	+0.73	+0.22
1964	−4	−18	14	−0.12	+1.18	+0.76

in meters per second at 30 millibars (approximately 23 km is altitude).[2] The R_s values in the fifth column are standard deviations from the mean rainfall for the western Sahel region. The values for R_g in the sixth column are standard deviations from the mean rainfall for the Gulf of Guinea. The dependent variable in the seventh column is the April to October rainfall in the western Sahel region based on recordings from 20 stations in the region.

For the multivariate data listed in Table 7.8, the LAD regression coefficients are

$$\tilde{\beta}_1 = -0.0021\,, \quad \tilde{\beta}_2 = -0.0364\,, \quad \tilde{\beta}_3 = -0.0325\,,$$

$$\tilde{\beta}_4 = +0.5328\,, \quad \text{and} \quad \tilde{\beta}_5 = +0.5215\,,$$

$\delta = 0.3439$, $\mu_\delta = 0.4756$, and the LAD chance-corrected measure of agreement between the observed y values and the predicted $\tilde{y}$ values is

$$\Re = 1 - \frac{\delta}{\mu_\delta} = 1 - \frac{0.3439}{0.4756} = +0.2768.$$

Even with a small sample of observations such as this, there are

$$M = N! = 15! = 1{,}307{,}674{,}368{,}000$$

[2]For comparison, the top of Mount Everest is approximately 8.85 km with a pressure of about 300 millibars.

possible, equally-likely arrangements of the observed to be considered, far too many for an exact permutation analysis. Based on $L = 1{,}000{,}000$ random arrangements of the observed data, the Monte Carlo resampling probability value of $\Re = +0.2768$ is

$$P\left(\Re \geq \Re_{\mathrm{o}} | H_0\right) = \frac{\text{number of } \Re \text{ values } \geq \Re_{\mathrm{o}}}{L} = \frac{42{,}279}{1{,}000{,}000} = 0.0423 \ ,$$

where $\Re_{\mathrm{o}}$ denotes the observed value of $\Re$.

7.3 LAD Multivariate Multiple Regression

An extension of LAD multiple linear regression to include multiple response variables, coupled with multiple predictor variables, is developed in this section [36, 37]. The extension was prompted by a multivariate Least Sum of Euclidean Distances (LSED) algorithm developed by Kaufman, Taylor, Mielke, and Berry in 2002 [24].

Consider the multivariate multiple linear regression model given by

$$y_{ik} = \sum_{j=1}^{m} x_{ij} \beta_{jk} + e_{ik}$$

for $i = 1, \ldots, N$ and $k = 1, \ldots, r$, where y_{ik} represents the ith of N measurements for the kth of r response variables, possibly affected by a treatment; x_{ij} is the jth of m covariates associated with the ith response, where $x_{i1} = 1$ if the model includes an intercept; β_{jk} denotes the jth of m regression parameters for the kth of r response variables; and e_{ik} designates the error associated with the ith of N measurements for the k of r response variables.

If estimates of β_{jk} that minimize

$$\sum_{i=1}^{N} \left(\sum_{k=1}^{r} e_{ik}^2 \right)^{1/2}$$

are denoted by $\tilde{\beta}_{jk}$ for $j = 1, \ldots, m$ and $k = 1, \ldots, r$, then the N r-dimensional residuals of the LSED multivariate multiple linear regression model are given by

$$e_{ik} = y_{ik} - \sum_{j=1}^{m} x_{ij} \tilde{\beta}_{jk}$$

for $i = 1, \ldots, N$ and $k = 1, \ldots, r$.

Let the N r-dimensional residuals, $e_{i1}, \ldots, e_{ir}$ for $i = 1, \ldots, N$, obtained from a LSED multivariate multiple linear regression model, be partitioned into g

treatment groups of sizes $n_1, \ldots, n_g$, where $n_i \geq 2$ for $i = 1, \ldots, g$ and

$$N = \sum_{i=1}^{g} n_i \ .$$

The analysis of the multivariate multiple regression residuals depends on test statistic

$$\delta = \sum_{i=1}^{g} C_i \xi_i \ , \tag{7.3}$$

where $C_i = n_i / N$ is a positive weight for the ith of g treatment groups and ξ_i is the average pairwise Euclidean distance among the n_i r-dimensional residuals in the ith of g treatment groups defined by

$$\xi_i = \binom{n_i}{2}^{-1} \sum_{k=1}^{N-1} \sum_{l=k+1}^{N} \left[\sum_{j=1}^{r} \left(e_{kj} - e_{lj} \right)^2 \right]^{1/2} \Psi_{ki} \Psi_{li} \ , \tag{7.4}$$

where

$$\Psi_{ki} = \begin{cases} 1 & \text{if } (e_{k1}, \ldots, e_{kr}) \text{ is in the } i\text{th treatment group} \ , \\ 0 & \text{otherwise} \ . \end{cases}$$

The null hypothesis specifies that each of the

$$M = \frac{N!}{\prod_{i=1}^{g} n_i!}$$

possible allocations of the N r-dimensional residuals to the g treatment groups is equally-likely. Under the null hypothesis, an exact probability value associated with the observed value of δ, δ_o, is given by

$$P(\delta \leq \delta_o | H_0) = \frac{\text{number of } \delta \text{ values } \leq \delta_o}{M} \ .$$

As with LAD univariate multiple regression models, the criterion for fitting LSED multivariate multiple regression models based on δ is the chance-corrected measure of effect size between the observed and predicted response measurement values given by

$$\Re = 1 - \frac{\delta}{\mu_\delta} \ , \tag{7.5}$$

where μ_δ is the expected value of δ over the $N!$ possible pairings under the null hypothesis, given by

$$\mu_\delta = \frac{1}{M} \sum_{i=1}^{M} \delta_i \ . \tag{7.6}$$

Note that $\Re = 1$ implies perfect agreement between the observed and model-predicted response vectors and the expected value of $\Re$ is 0 under the null hypothesis, i.e., chance-corrected.

7.3.1 Example of Multivariate Multiple Regression

To illustrate a multivariate LSED multiple regression analysis, consider an unbalanced two-way randomized-block experimental design in which $N = 16$ subjects are tested over $a = 3$ levels of Factor A, the experiment is repeated $b = 2$ times for Factor B, and there are $r = 2$ response measurement scores for each subject. The data are listed in Table 7.9. The design is intentionally kept small to illustrate the multivariate multiple regression procedure.

Analysis of Factor *A*

A design matrix of dummy codes (0, 1) for a regression analysis of Factor A is given in Table 7.10, where the first column of 1 values provides for an intercept, the next column contains the dummy codes for Factor B, and the third and fourth columns contain the bivariate response measurement scores listed according to the original random assignment of the $N = 16$ subjects to the $a = 3$ levels of Factor A, with the first $n_{A_1} = 5$ scores, the next $n_{A_2} = 7$ scores, and the last $n_{A_3} = 4$ scores associated with the $a = 3$ levels of Factor A, respectively. The analysis of

Table 7.9 Example data for a two-way randomized-block design with $a = 3$ blocks, $b = 2$ treatments, and $N = 16$ subjects

	Factor A		
Factor B	A_1	A_2	A_3
B_1	(49, 102)	(63, 84)	(45, 107)
		(60, 89)	(50, 100)
			(42, 111)
			(46, 104)
B_2	(48, 103)	(27, 114)	
	(58, 94)	(66, 83)	
	(51, 100)	(74, 79)	
	(55, 97)	(69, 88)	
		(71, 82)	

Table 7.10 Example design matrix and bivariate response measurement scores for a multivariate LSED multiple regression analysis of Factor A with $N = 16$ subjects

Matrix		Scores	
1	1	49	102
1	0	48	103
1	0	58	94
1	0	51	100
1	0	55	97
1	1	63	84
1	1	60	89
1	0	27	114
1	0	66	83
1	0	74	79
1	0	69	88
1	0	71	82
1	1	45	107
1	1	50	100
1	1	42	111
1	1	46	104

the data listed in Table 7.10 examines the $N = 16$ regression residuals for possible differences among the $a = 3$ treatment levels of Factor A; consequently, no dummy codes are provided for Factor A as this information is implicit in the ordering of the $a = 3$ levels of Factor A in the last two columns of Table 7.10.

Because there are only

$$M = \frac{N!}{\prod_{i=1}^{a} n_{A_i}!} = \frac{16!}{5!\ 7!\ 4!} = 1{,}441{,}440$$

possible, equally-likely arrangements of the $N = 16$ bivariate response measurement scores listed in Table 7.10, an exact permutation analysis is feasible. The analysis of the $N = 16$ LAD regression residuals calculated on the bivariate response measurement scores for Factor A in Table 7.10 yields estimated LAD regression coefficients of

$$\tilde{\beta}_{1,1} = +58.00\,, \quad \tilde{\beta}_{2,1} = -9.00\,, \quad \tilde{\beta}_{1,2} = +94.00\,, \quad \text{and} \quad \tilde{\beta}_{2,2} = +8.00$$

for Factor A. Table 7.11 lists the observed y_{ik} values, LAD-predicted $\tilde{y}_{ik}$ values, and residual e_{ik} values for $i = 1, \ldots, 16$ subjects and $k = 1, 2$ response variables.

Following Eq. (7.4) on p. 393 and employing ordinary Euclidean distance between residuals, the $N = 16$ LAD regression residuals listed in Table 7.11 yield $a = 3$ average distance-function values of

$$\xi_{A_1} = 7.2294\,, \quad \xi_{A_2} = 20.0289\,, \quad \text{and} \quad \xi_{A_3} = 7.3475\,.$$

Table 7.11 Observed, predicted, and residual values for a multivariate LSED multiple regression analysis of Factor A with $N = 16$ subjects

y_{i1}	y_{i2}	$\tilde{y}_{i1}$	$\tilde{y}_{i2}$	e_{i1}	e_{i2}
49	102	49.00	102.00	0.00	0.00
48	103	58.00	94.00	−10.00	+9.00
58	94	58.00	94.00	0.00	0.00
51	100	58.00	94.00	−7.00	+6.00
55	97	58.00	94.00	−3.00	+3.00
63	84	49.00	102.00	+14.00	−18.00
60	89	49.00	102.00	+11.00	−13.00
27	114	58.00	94.00	−31.00	+20.00
66	83	58.00	94.00	+8.00	−11.00
74	79	58.00	94.00	+16.00	−15.00
69	88	58.00	94.00	+11.00	−6.00
71	82	58.00	94.00	+13.00	−12.00
45	107	49.00	102.00	−4.00	+5.00
50	100	49.00	102.00	+1.00	−2.00
42	111	49.00	102.00	−7.00	+9.00
46	104	49.00	102.00	−3.00	+2.00

Following Eq. (7.3) on p. 393, the observed value of test statistic δ calculated on the $N = 16$ LAD regression residuals listed in Table 7.11 with treatment group weights

$$C_j = \frac{n_{A_j}}{N} \qquad \text{for } j = 1, 2, 3$$

is

$$\delta_A = \sum_{j=1}^{a} C_j \xi_j = \frac{1}{16}\big[(5)(7.2294) + (7)(20.0289) + (4)(7.3475)\big] = 12.8587 \, .$$

If all M arrangements of the $N = 16$ observed LAD regression residuals listed in Table 7.11 occur with equal chance, the exact probability value of $\delta_A = 12.8587$ computed on the $M = 1{,}441{,}440$ possible arrangements of the observed LAD regression residuals with $n_{A_1} = 5$, $n_{A_2} = 7$, and $n_{A_3} = 4$ preserved for each arrangement is

$$P(\delta \le \delta_A | H_0) = \frac{\text{number of } \delta \text{ values } \le \delta_A}{M} = \frac{6{,}676}{1{,}441{,}440} = 0.0046 \, .$$

Following Eq. (7.6) on p. 394, the exact expected value of the $M = 1{,}441{,}440$ δ values is

$$\mu_\delta = \frac{1}{M} \sum_{i=1}^{M} \delta_i = \frac{26{,}092{,}946.8800}{1{,}441{,}440} = 18.1020$$

and, following Eq. (7.5) on p. 393, the observed chance-corrected measure of effect size for the y_i and $\tilde{y}_i$ values, $i = 1, \ldots, N$, is

$$\Re_A = 1 - \frac{\delta_A}{\mu_\delta} = 1 - \frac{12.8587}{18.1020} = +0.2897 \,,$$

indicating approximately 29% agreement between the observed and predicted values above that expected by chance.

Analysis of Factor *B*

A design matrix of dummy codes (0, 1) for a regression analysis of Factor B is given in Table 7.12, where the first column of 1 values provides for an intercept, the next two columns contain the dummy codes for Factor A, and the fourth and fifth columns contain the bivariate response measurement scores listed according to the original random assignment of the $N = 16$ subjects to the $b = 2$ levels of Factor B, with the first $n_{B_1} = 7$ scores and the last $n_{B_2} = 9$ scores associated with the $b = 2$ levels of Factor B, respectively. The analysis of the data listed in Table 7.12 examines the $N = 16$ regression residuals for possible differences between the $b = 2$ treatment levels of Factor B; consequently, no dummy codes are provided for Factor B as this information is implicit in the ordering of the $b = 2$ levels of Factor B in the last two columns of Table 7.12.

Table 7.12 Example design matrix and bivariate response measurement scores for a multivariate LSED multiple regression analysis of Factor B with $N = 16$ subjects

Matrix			Scores	
1	1	0	49	102
1	0	1	63	84
1	0	1	60	89
1	0	0	45	107
1	0	0	50	100
1	1	0	42	111
1	1	0	46	104
1	0	0	48	103
1	0	0	58	94
1	0	0	51	100
1	0	0	55	97
1	0	1	27	114
1	1	1	66	83
1	1	1	74	79
1	1	1	69	88
1	1	1	71	82

Because there are only

$$M = \frac{N!}{\prod_{i=1}^{b} n_{B_i}!} = \frac{16!}{7!\,9!} = 11{,}440$$

possible, equally-likely arrangements of the $N = 16$ response measurement scores listed in Table 7.12, an exact permutation analysis is feasible. The analysis of the $N = 16$ LAD regression residuals calculated on the bivariate response measurement scores for Factor B in Table 7.12 yields estimated LAD regression coefficients of

$$\tilde{\beta}_{1,1} = +46.00\,, \quad \tilde{\beta}_{2,1} = +5.00\,, \quad \tilde{\beta}_{3,1} = +20.00\,, \quad \tilde{\beta}_{1,2} = +104.00\,,$$
$$\tilde{\beta}_{2,2} = -4.00\,, \quad \text{and} \quad \tilde{\beta}_{3,2} = -20.00$$

for Factor B. Table 7.13 lists the observed y_{ik} values, LAD-predicted $\tilde{y}_{ik}$ values, and residual e_{ik} values for $i = 1, \ldots, 16$ subjects and $k = 1, 2$ response variables.

Following Eq. (7.4) on p. 393 and employing ordinary Euclidean distance between residuals, the $N = 16$ LAD regression residuals listed in Table 7.13 yield $b = 2$ average distance-function values of

$$\xi_{B_1} = 6.0229 \quad \text{and} \quad \xi_{B_2} = 16.7440\,.$$

Table 7.13 Observed, predicted, and residual values for a multivariate LSED multiple regression analysis of Factor A with $N = 16$ subjects

y_{i1}	y_{i2}	$\tilde{y}_{i1}$	$\tilde{y}_{i2}$	e_{i1}	e_{i2}
49	102	51.00	100.00	−2.00	+2.00
63	84	66.00	84.00	−3.00	0.00
60	89	66.00	84.00	−6.00	+5.00
45	107	46.00	104.00	−1.00	+3.00
50	100	46.00	104.00	+4.00	−4.00
42	111	46.00	104.00	−4.00	+7.00
46	104	46.00	104.00	0.00	0.00
48	103	51.00	100.00	−3.00	+3.00
58	94	51.00	100.00	+7.00	−6.00
51	100	51.00	100.00	0.00	0.00
55	97	51.00	100.00	+4.00	−3.00
27	114	66.00	84.00	−39.00	+30.00
66	83	66.00	84.00	0.00	−1.00
74	79	66.00	84.00	−8.00	−5.00
69	88	66.00	84.00	+3.00	+4.00
71	82	66.00	84.00	+5.00	−2.00

Following Eq. (7.3) on p. 393, the observed value of test statistic δ calculated on the $N = 16$ LAD regression residuals listed in Table 7.13 with treatment group weights

$$C_i = \frac{n_{B_i}}{N} \qquad \text{for } i = 1, 2 \,,$$

is

$$\delta_B = \sum_{i=1}^{b} C_i \xi_i = \frac{1}{16}\big[(7)(6.0229) + (9)(16.7440)\big] = 12.0535 \,.$$

If all M arrangements of the $N = 16$ observed LAD regression residuals listed in Table 7.13 occur with equal chance, the exact probability value of $\delta_B = 12.0535$ computed on the $M = 11{,}440$ possible arrangements of the observed LAD regression residuals with $n_{B_1} = 7$ and $n_{B_2} = 9$ preserved for each arrangement is

$$P(\delta \le \delta_B | H_0) = \frac{\text{number of } \delta \text{ values } \le \delta_B}{M} = \frac{2{,}090}{11{,}440} = 0.1827 \,.$$

Following Eq. (7.6) on p. 394, the exact expected value of the $M = 11{,}440$ δ values is

$$\mu_\delta = \frac{1}{M}\sum_{i=1}^{M} \delta_i = \frac{140{,}623.9120}{11{,}440} = 12.2923$$

and, following Eq. (7.5) on p. 393, the observed chance-corrected measure of effect size for the y_i and $\tilde{y}_i$ values, $i = 1, \ldots, N$, is

$$\Re_B = 1 - \frac{\delta_B}{\mu_\delta} = 1 - \frac{12.0535}{12.2923} = +0.0194 \,,$$

indicating approximately 2% agreement between the observed and predicted values above that expected by chance.

For another example of LAD multiple multivariate example, see an informative and widely cited article by Endler and Mielke on "Comparing entire colour patterns as birds see them" in *Biological Journal of the Linnean Society* [11].

7.4 Comparison of OLS and LAD Linear Regression

In this section, OLS and LAD linear regression analyses are illustrated and compared on two example data sets—one with $p = 2$ predictors and no extreme

Table 7.14 Example multivariate correlation data on $N = 12$ families with $p = 2$ predictors

Family	x_1	x_2	y
A	1	12	1
B	1	14	2
C	1	16	3
D	1	16	5
E	2	18	3
F	2	16	1
G	3	12	5
H	3	12	0
I	4	10	6
J	4	12	3
K	5	10	7
L	5	16	4

values and one with $p = 2$ predictors and a single extreme value.[3] Consider first the small example data set with $p = 2$ predictors listed in Table 7.14 where variable y is Hours of Housework done by husbands per week, variable x_1 is Number of Children, and variable x_2 is husband's Years of Education for $N = 12$ families.

7.4.1 Ordinary Least Squares (OLS) Analysis

For the multivariate data listed in Table 7.14, the unstandardized OLS regression coefficients are

$$\hat{\beta}_1 = +0.6356 \quad \text{and} \quad \hat{\beta}_2 = -0.0649 \,,$$

and the observed squared OLS multiple correlation coefficient is $R^2_{\text{o}} = 0.2539$. Based on $L = 1{,}000{,}000$ random arrangements of the observed data, the Monte Carlo resampling probability value of $R^2_{\text{o}} = 0.2539$ is

$$P\big(R^2 \geq R^2_{\text{o}} | H_0\big) = \frac{\text{number of } R^2 \text{ values } \geq R^2_{\text{o}}}{L} = \frac{268{,}026}{1{,}000{,}000} = 0.2680 \,,$$

where R^2_{o} denotes the observed value of R^2. For comparison, the exact probability value of $R^2_{\text{o}} = 0.2539$ based on $M = N! = 12! = 479{,}001{,}600$ possible arrangements of the data listed in Table 7.14 is $P = 0.2681$.

[3] For real-life applications and comparisons of OLS and LAD regression applied to meteorological forecasting, see two articles in *Weather and Forecasting* by Mielke, Berry, Landsea, and Gray [39, 40].

7.4.2 Least Absolute Deviation (LAD) Analysis

For the multivariate data listed in Table 7.14, the LAD regression coefficients are

$$\tilde{\beta}_1 = +0.4138 \quad \text{and} \quad \tilde{\beta}_2 = +0.1207 ,$$

$\delta = 1.5000$, $\mu_\delta = 1.8084$, and the LAD chance-corrected measure of agreement between the observed y values and the predicted $\tilde{y}$ values is

$$\Re_o = 1 - \frac{\delta}{\mu_\delta} = 1 - \frac{1.5000}{1.8084} = +0.1706 .$$

Based on $L = 1{,}000{,}000$ random arrangements of the observed data, the Monte Carlo resampling probability value of $\Re = +0.1706$ is

$$P\left(\Re \geq \Re_o | H_0\right) = \frac{\text{number of } \Re \text{ values } \geq \Re_o}{L} = \frac{19{,}176}{1{,}000{,}000} = 0.0192 ,$$

where $\Re_o$ denotes the observed value of $\Re$. For comparison, the exact probability value of $\Re_o = +0.1706$ based on $M = N! = 12! = 479{,}001{,}600$ possible arrangements of the data listed in Table 7.14 is $P = 0.0221$.

Now, suppose that the husband in family "L" was a stay-at-home house-husband and instead of contributing just four hours of housework per week, he actually contributed 40 hours, as in Table 7.15.

Table 7.15 Example multivariate correlation data on $N = 12$ families with $p = 2$ predictors, where the husband in Family L contributed 40 hours of housework per week

Family	x_1	x_2	y
A	1	12	1
B	1	14	2
C	1	16	3
D	1	16	5
E	2	18	3
F	2	16	1
G	3	12	5
H	3	12	0
I	4	10	6
J	4	12	3
K	5	10	7
L	5	16	40

7.4.3 Ordinary Least Squares (OLS) Analysis

For the multivariate data listed in Table 7.15, the unstandardized OLS regression coefficients are

$$\hat{\beta}_1 = +5.7492 \quad \text{and} \quad \hat{\beta}_2 = +2.3896 ,$$

and the observed squared OLS multiple correlation coefficient is $R^2_{\text{o}} = 0.5786$. Based on $L = 1{,}000{,}000$ random arrangements of the observed data, the Monte Carlo resampling probability value of $R^2_{\text{o}} = 0.5786$ is

$$P\left(R^2 \geq R^2_{\text{o}} | H_0\right) = \frac{\text{number of } R^2 \text{ values } \geq R^2_{\text{o}}}{L} = \frac{15{,}215}{1{,}000{,}000} = 0.0152 ,$$

where R^2_{o} denotes the observed value of R^2. For comparison, the exact probability value of $R^2_{\text{o}} = 0.5786$ based on $M = N! = 12! = 479{,}001{,}600$ possible arrangements of the data listed in Table 7.15 is $P = 0.0153$.

7.4.4 Least Absolute Deviation (LAD) Analysis

For the multivariate data listed in Table 7.15, the LAD regression coefficients are

$$\tilde{\beta}_1 = +1.3000 \quad \text{and} \quad \tilde{\beta}_2 = +0.0500 ,$$

$\delta_{\text{o}} = 4.0333$, $\mu_\delta = 5.2194$, and the LAD chance-corrected measure of agreement between the observed y values and the predicted $\tilde{y}$ values is

$$\Re_{\text{o}} = 1 - \frac{\delta_{\text{o}}}{\mu_\delta} = 1 - \frac{4.0333}{5.2194} = +0.2272 .$$

Based on $L = 1{,}000{,}000$ random arrangements of the observed data, the Monte Carlo resampling probability value of $\Re_{\text{o}} = +0.2272$ is

$$P\left(\Re \geq \Re_{\text{o}} | H_0\right) = \frac{\text{number of } \Re \text{ values } \geq \Re_{\text{o}}}{L} = \frac{4{,}517}{1{,}000{,}000} = 0.4571 \times 10^{-2} ,$$

where $\Re_{\text{o}}$ denotes the observed value of $\Re$. For comparison, the exact probability value of $\Re_{\text{o}} = +0.2272$ based on $M = N! = 12! = 479{,}001{,}600$ possible arrangements of the data listed in Table 7.14 is $P = 0.5630 \times 10^{-2}$.

The results of the comparison of OLS and LAD analyses with 4 and 40 hours of housework by the husband in family "L" are summarized in Table 7.16. The value

Table 7.16 Comparison of OLS and LAD analyses for the data given in Table 7.14 with 4 hours of housework for the husband in family L and the data given in Table 7.15 with 40 hours of housework for the husband in family L

	OLS analysis		LAD analysis	
Hours	R^2	Probability	$\Re$	Probability
4	0.2539	0.2680	0.1706	0.0192
40	0.5786	0.0152	0.2272	0.0046
$\lvert\Delta\rvert$	0.3247	0.2528	0.0566	0.0146

of 40 hours of housework by the husband in family "L" is, by any definition, an extreme value. It is six times the mean of $\bar{y} = 6.3333$ and three standard deviations above the mean. It is readily apparent that the extreme value of 40 hours had a profound impact on the results of the OLS analysis. The OLS multiple correlation coefficient more than doubled from $R_o^2 = 0.2539$ to $R_o^2 = 0.5786$, a difference of $R^2 = 0.3247$, and the corresponding probability value decreased from $P = 0.2680$ to $P = 0.0152$, a difference of $P = 0.2528$. The impact of 40 hours of housework on the LAD analysis is more modest with the LAD chance-corrected measure of agreement increasing only slightly from $\Re_o = 0.1706$ to $\Re_o = 0.2272$, a difference of $\Re = 0.0566$, and the probability value decreasing from $P = 0.0192$ to $P = 0.0046$, a difference of only $P = 0.0146$.

7.5 Fisher's r_{xy} to z Transformation

In order to attach a probability statement to inferences about the Pearson product-moment correlation coefficient, it is necessary to know the sampling distribution of a statistic that relates the sample correlation coefficient, r_{xy}, to the population parameter, ρ_{xy}. Because $-1.0 \leq r_{xy} \leq +1.0$, the sampling distribution of statistic r_{xy} is asymmetric whenever $\rho_{xy} \neq 0.0$.[4] Given two random variables that follow the bivariate normal distribution with population parameter ρ_{xy}, the sampling distribution of statistic r_{xy} approaches normality as the sample size increases; however, it converges very slowly for $|\rho_{xy}| \geq 0.6$, even with samples as large as $N = 400$ [7, p. xxxiii]. Fisher [13, 14] obtained the basic distribution of r_{xy} and showed that, when bivariate normality is assumed, a logarithmic transformation of r_{xy} (henceforth referred to as the Fisher z transform),

$$z = \frac{1}{2} \ln \left(\frac{1 + r_{xy}}{1 - r_{xy}} \right) = \tanh^{-1}(r_{xy}) \ ,$$

[4] It is probably safe to assume that in any actual research situation, the population correlation coefficient is always not equal to zero.

becomes normally distributed with a mean of approximately

$$\frac{1}{2}\ln\left(\frac{1+\rho_{xy}}{1-\rho_{xy}}\right) = \tanh^{-1}(\rho_{xy})$$

and the standard error approaches

$$\frac{1}{\sqrt{N-3}}$$

as $N \to \infty$.

The Fisher r_{xy} to z transform is presented in most textbooks and is available in a wide array of statistical software packages. In this section, the precision and accuracy of the Fisher z transform are examined for a variety of bivariate distributions, sample sizes, and values of ρ_{xy} [5]. If $\rho_{xy} \neq 0.0$ and the distribution is not bivariate normal, then the desired properties of the Fisher z transform generally fail.

There are two general applications of the Fisher z transform. The first application comprises the computation of the confidence limits for ρ_{xy} and the second involves the testing of hypotheses about specified values of $\rho_{xy} \neq 0.0$. The second application is more tractable than the first application as a hypothesized value of ρ_{xy} is available. The next part of this section describes the bivariate distributions to be examined, followed by an exploration of confidence intervals and an examination of hypothesis testing. The last part of the section provides some general conclusions about the propriety of uncritically using the Fisher z transform in actual research.

7.5.1 Distributions

Seven bivariate distributions are utilized to test the Fisher z transform. In addition, two related methods by Gayen [17] and Jeyaratnam [22] are also examined. The Gayen and Jeyaratnam techniques are characterized by simplicity, accuracy, and ease of use. For other interesting approaches, see David [7]; Hotelling [21]; Kraemer [25]; Liu, Woodward, and Bonett [28]; Mudholkar and Chaubey [41]; Pillai [45]; Ruben [48]; and Samiuddin [49].

Normal Distribution

The density function of the standardized normal, $N(0, 1)$, distribution is given by

$$f(x) = (2\pi)^{-1/2}\exp(-x^2/2)\ .$$

Generalized Logistic Distribution

The density function of the generalized logistic (GL) distribution is given by

$$f(x) = \left[\exp(\theta x)/\theta\right]^{1/\theta}\left[1 + \exp(\theta x)/\theta\right]^{-(\theta+1)/\theta}$$

for $\theta > 0$ [34]. The generalized logistic distribution is positively skewed for $\theta < 1$ and negatively skewed for $\theta > 1$. When $\theta = 1.0$, $GL(\theta)$ is a logistic distribution that closely resembles the normal distribution, with somewhat lighter tails. When $\theta = 0.10$, $GL(\theta)$ is a generalized logistic distribution with positive skewness. When $\theta = 0.01$, $GL(\theta)$ is a generalized logistic distribution with even greater positive skewness.

Symmetric Kappa Distribution

The density function of the symmetric kappa (SK) distribution is given by

$$f(x) = 0.5\lambda^{-1/\lambda}\left(1 + |x|^{\lambda}/\lambda\right)^{-(\lambda+1)/\lambda}$$

for $\lambda > 0$ [34, 35]. The shape of the symmetric kappa distribution ranges from an exceedingly heavy-tailed distribution as λ approaches zero to a uniform distribution as λ goes to infinity. When $\lambda = 2$, $SK(\lambda)$ is a peaked, heavy-tailed distribution, identical to Student's t distribution with 2 degrees of freedom. Thus, the variance of $SK(2)$ does not exist. When $\lambda = 3$, $SK(\lambda)$ is also a heavy-tailed distribution, but the variance does exist. When $\lambda = 25$, $SK(\lambda)$ is a loaf-shaped distribution resembling a uniform distribution with the addition of very light tails. These distributions provide a variety of populations from which to sample and evaluate the Fisher z transformation and the Gayen [17] and Jeyaratnam [22] modifications.

Seven bivariate correlated distributions were constructed in the following manner. Let x and y be independent identically distributed univariate random variables from each of seven univariate distributions, i.e., $N(0, 1)$, $GL(1.0)$, $GL(0.1)$, $GL(0.01)$, $SK(2)$, $SK(3)$, and $SK(25)$, and define the correlated random variables U_1 and U_2 of each bivariate distribution by

$$U_1 = x(1 - \rho_{xy}^2)^{1/2} + \rho_{xy} y$$

and $U_2 = y$, where ρ_{xy} is the desired Pearson product-moment correlation coefficient of random variables U_1 and U_2. Then a Monte Carlo procedure obtains random samples, corresponding to x and y, from the normal, generalized logistic, and symmetric kappa distributions.

7.5.2 Confidence Intervals

In this section, Monte Carlo confidence intervals are based on the seven distributions: $N(0, 1)$, $GL(1.0)$, $GL(0.1)$, $GL(0.01)$, $SK(2)$, $SK(3)$, and $SK(25)$. Each simulation is based on $L = 1{,}000{,}000$ bivariate random samples, U_1 and U_2, of size $N = 10$, 20, 40, and 80 for $\rho_{xy} = 0.00$, $+0.40$, $+0.60$, and $+0.80$ with $1 - \alpha = 0.90$, 0.95, and 0.99. Confidence intervals obtained from two methods are considered. The first confidence interval is based on the Fisher z transform and is defined by

$$\tanh\left[\tanh^{-1}(r_{xy}) - \frac{z_{\alpha/2}}{\sqrt{N-3}}\right] \le \rho_{xy} \le \tanh\left[\tanh^{-1}(r_{xy}) + \frac{z_{\alpha/2}}{\sqrt{N-3}}\right],$$

where $z_{\alpha/2}$ is the upper 0.50α probability point of the $N(0, 1)$ distribution. The second confidence interval is based on a method proposed by Jeyaratnam [22] and is defined by

$$\frac{r_{xy} - w}{1 - r_{xy}w} \le \rho_{xy} \le \frac{r_{xy} + w}{1 + r_{xy}w},$$

where

$$w = \frac{\left(t_{\alpha/2,N-2}\right)/\sqrt{N-2}}{\left[1 + \left(t_{\alpha/2,N-2}\right)^2/\sqrt{N-2}\right]^{1/2}}$$

and $t_{\alpha/2,N-2}$ is the upper 0.50α probability point of Student's t distribution with $N - 2$ degrees of freedom.

The results of the Monte Carlo analyses are summarized in Tables 7.17, 7.18, 7.19, 7.20, 7.21, 7.22, 7.23, which contain simulated containment probability values for the seven bivariate distributions with specified nominal values of $1 - \alpha$ (0.90, 0.95, 0.99), ρ_{xy} (0.00, +0.40, +0.60, +0.80), and N (10, 20, 40, 80) for the Fisher (F) and Jeyaratnam (J) confidence intervals. Table 7.17 analyzes data obtained from the $N(0, 1)$ distribution; Tables 7.18, 7.19, and 7.20 analyze data obtained from the generalized logistic distribution with $\theta = 1.0$, 0.1, and 0.01, respectively; and Tables 7.21, 7.22, and 7.23 analyze data obtained from the symmetric kappa distribution with $\lambda = 2$, 3, and 25, respectively.

In each of the seven tables, the Monte Carlo containment probability values for a $1 - \alpha$ confidence interval based on the Fisher z transform and a $1 - \alpha$ confidence interval based on the Jeyaratnam technique were obtained from the same $L = 1{,}000{,}000$ bivariate random samples of size N drawn with replacement from the designated bivariate distribution characterized by the specified population correlation ρ_{xy}. If the Fisher and Jeyaratnam transforms are appropriate for the

Table 7.17 Containment probability values for a bivariate $N(0, 1)$ distribution with Fisher (F) and Jeyaratnam (J) $1 - \alpha$ correlation confidence intervals

		$\rho_{xy} = 0.00$		$\rho_{xy} = +0.40$		$\rho_{xy} = +0.60$		$\rho_{xy} = +0.80$	
$1 - \alpha$	N	F	J	F	J	F	J	F	J
0.90	10	0.9014	0.8992	0.9026	0.9004	0.9037	0.9015	0.9048	0.9025
	20	0.9012	0.9005	0.9015	0.9008	0.9009	0.9002	0.9020	0.9014
	40	0.9004	0.9001	0.9012	0.9009	0.9009	0.9006	0.9011	0.9009
	80	0.9002	0.9001	0.9000	0.9000	0.9006	0.9005	0.9008	0.9007
0.95	10	0.9491	0.9501	0.9490	0.9501	0.9497	0.9508	0.9516	0.9516
	20	0.9495	0.9502	0.9493	0.9501	0.9500	0.9507	0.9500	0.9507
	40	0.9495	0.9499	0.9497	0.9501	0.9493	0.9497	0.9502	0.9506
	80	0.9595	0.9498	0.9497	0.9499	0.9501	0.9503	0.9498	0.9500
0.99	10	0.9875	0.9900	0.9877	0.9900	0.9877	0.9901	0.9880	0.9904
	20	0.9889	0.9900	0.9888	0.9900	0.9890	0.9901	0.9891	0.9902
	40	0.9893	0.9899	0.9896	0.9901	0.9894	0.9900	0.9895	0.9901
	80	0.9896	0.9899	0.9897	0.9900	0.9897	0.9900	0.9897	0.9900

Table 7.18 Containment probability values for a bivariate $GL(1.0)$ distribution with Fisher (F) and Jeyaratnam (J) $1 - \alpha$ correlation confidence intervals

		$\rho_{xy} = 0.00$		$\rho_{xy} = +0.40$		$\rho_{xy} = +0.60$		$\rho_{xy} = +0.80$	
$1 - \alpha$	N	F	J	F	J	F	J	F	J
0.90	10	0.9011	0.8990	0.8930	0.8907	0.8833	0.8809	0.8710	0.8684
	20	0.9009	0.9002	0.8894	0.8886	0.8742	0.8734	0.8565	0.8557
	40	0.9007	0.9004	0.8873	0.8871	0.8701	0.8698	0.8484	0.8481
	80	0.9005	0.9004	0.8851	0.8850	0.8677	0.8676	0.8438	0.8437
0.95	10	0.9485	0.9496	0.9425	0.9437	0.9359	0.9372	0.9273	0.9287
	20	0.9491	0.9498	0.9407	0.9415	0.9313	0.9322	0.9170	0.9181
	40	0.9491	0.9496	0.9402	0.9406	0.9274	0.9279	0.9116	0.9121
	80	0.9497	0.9499	0.9394	0.9396	0.9266	0.9269	0.9082	0.9085
0.99	10	0.9873	0.9897	0.9852	0.9880	0.9827	0.9858	0.9794	0.9832
	20	0.9886	0.9897	0.9855	0.9870	0.9821	0.9838	0.9764	0.9785
	40	0.9891	0.9897	0.9861	0.9867	0.9815	0.9823	0.9744	0.9755
	80	0.9895	0.9898	0.9860	0.9864	0.9808	0.9812	0.9729	0.9735

simulated data, the containment probability values should agree with the nominal $1 - \alpha$ values.

Some general observations can be made about the Monte Carlo results contained in Tables 7.17 through 7.23. First, in each of the tables there is little difference between the Fisher and Jeyaratnam Monte Carlo containment probability values and both techniques provide values close to the nominal $1 - \alpha$ values for the $N(0, 1)$ distribution analyzed in Table 7.17 with any value of ρ_{xy} and for any of the other distributions analyzed in Tables 7.18 through 7.23 when $\rho_{xy} = 0.00$. Second, for the skewed and heavy-tailed distributions, i.e., $GL(0.1)$, $GL(0.01)$, $SK(2)$, and

Table 7.19 Containment probability values for a bivariate $GL(0.1)$ distribution with Fisher (F) and Jeyaratnam (J) $1 - \alpha$ correlation confidence intervals

		$\rho_{xy} = 0.00$		$\rho_{xy} = +0.40$		$\rho_{xy} = +0.60$		$\rho_{xy} = +0.80$	
$1 - \alpha$	N	F	J	F	J	F	J	F	J
0.90	10	0.9016	0.8995	0.8878	0.8854	0.8729	0.8704	0.8544	0.8516
	20	0.9013	0.9006	0.8821	0.8813	0.8593	0.8584	0.8321	0.8313
	40	0.9010	0.9007	0.8780	0.8777	0.8510	0.8507	0.8174	0.8170
	80	0.9006	0.9004	0.8760	0.8759	0.8459	0.8457	0.8081	0.8079
0.95	10	0.9486	0.9497	0.9389	0.9401	0.9281	0.9295	0.9150	0.9165
	20	0.9495	0.9502	0.9354	0.9362	0.9197	0.9206	0.8982	0.8993
	40	0.9495	0.9499	0.9335	0.9340	0.9136	0.9141	0.8871	0.8877
	80	0.9498	0.9500	0.9320	0.9323	0.9100	0.9102	0.8797	0.8800
0.99	10	0.9871	0.9895	0.9835	0.9865	0.9793	0.9830	0.9744	0.9787
	20	0.9882	0.9895	0.9833	0.9850	0.9770	0.9790	0.9674	0.9700
	40	0.9890	0.9895	0.9833	0.9841	0.9752	0.9763	0.9623	0.9637
	80	0.9895	0.9898	0.9828	0.9832	0.9737	0.9743	0.9585	0.9592

Table 7.20 Containment probability values for a bivariate $GL(0.01)$ distribution with Fisher (F) and Jeyaratnam (J) $1 - \alpha$ correlation confidence intervals

		$\rho_{xy} = 0.00$		$\rho_{xy} = +0.40$		$\rho_{xy} = +0.60$		$\rho_{xy} = +0.80$	
$1 - \alpha$	N	F	J	F	J	F	J	F	J
0.90	10	0.9019	0.8996	0.8860	0.8837	0.8693	0.8667	0.8485	0.8457
	20	0.9015	0.9008	0.8798	0.8790	0.8545	0.8537	0.8243	0.8234
	40	0.9012	0.9009	0.8754	0.8752	0.8454	0.8450	0.8084	0.8080
	80	0.9002	0.9001	0.8726	0.8724	0.8394	0.8393	0.7984	0.7982
0.95	10	0.9485	0.9496	0.9375	0.9388	0.9255	0.9269	0.9106	0.9121
	20	0.9496	0.9503	0.9337	0.9346	0.9160	0.9170	0.8921	0.8932
	40	0.9495	0.9499	0.9317	0.9321	0.9092	0.9097	0.8797	0.8803
	80	0.9500	0.9502	0.9296	0.9298	0.9055	0.9057	0.8713	0.8716
0.99	10	0.9869	0.9893	0.9829	0.9860	0.9782	0.9820	0.9725	0.9771
	20	0.9881	0.9893	0.9825	0.9842	0.9752	0.9774	0.9644	0.9671
	40	0.9889	0.9895	0.9825	0.9833	0.9732	0.9743	0.9584	0.9600
	80	0.9897	0.9897	0.9821	0.9825	0.9712	0.9718	0.9540	0.9548

$SK(3)$, with N held constant, the differences between the Monte Carlo containment probability values and the nominal $1 - \alpha$ values become greater as $|\rho_{xy}|$ increases. Third, the differences between the Monte Carlo containment probability values and the nominal $1 - \alpha$ values increase with increasing N and $|\rho_{xy}| > 0.00$ for all the distributions except $N(0, 1)$ and $SK(25)$. This is especially evident with the skewed and heavy-tailed distributions $GL(0.1)$, $GL(0.01)$, $SK(2)$, and $SK(3)$.

Table 7.21 Containment probability values for a bivariate $SK(2)$ distribution with Fisher (F) and Jeyaratnam (J) $1-\alpha$ correlation confidence intervals

		$\rho_{xy} = 0.00$		$\rho_{xy} = +0.40$		$\rho_{xy} = +0.60$		$\rho_{xy} = +0.80$	
$1-\alpha$	N	F	J	F	J	F	J	F	J
0.90	10	0.8961	0.8942	0.8054	0.8029	0.7487	0.7457	0.6806	0.6774
	20	0.9002	0.8996	0.7582	0.7573	0.6650	0.6641	0.5733	0.5723
	40	0.9050	0.9048	0.6968	0.6965	0.5755	0.5752	0.4784	0.4781
	80	0.9097	0.9096	0.6192	0.6191	0.4884	0.4883	0.3942	0.3941
0.95	10	0.9403	0.9413	0.8670	0.8687	0.8198	0.8217	0.7612	0.7634
	20	0.9415	0.9421	0.8257	0.8269	0.7442	0.7457	0.6522	0.6538
	40	0.9436	0.9439	0.7726	0.7732	0.6543	0.6551	0.5521	0.5528
	80	0.9461	0.9463	0.6982	0.6986	0.5630	0.5634	0.4599	0.4602
0.99	10	0.9797	0.9828	0.9357	0.9420	0.9068	0.9152	0.8697	0.8810
	20	0.9789	0.9803	0.9065	0.9102	0.8523	0.8577	0.7761	0.7829
	40	0.9788	0.9794	0.8694	0.8715	0.7748	0.7780	0.6733	0.6768
	80	0.9794	0.9797	0.8107	0.8121	0.6819	0.6835	0.5721	0.5738

Table 7.22 Containment probability values for a bivariate $SK(3)$ distribution with Fisher (F) and Jeyaratnam (J) $1-\alpha$ correlation confidence intervals

		$\rho_{xy} = 0.00$		$\rho_{xy} = +0.40$		$\rho_{xy} = +0.60$		$\rho_{xy} = +0.80$	
$1-\alpha$	N	F	J	F	J	F	J	F	J
0.90	10	0.9007	0.8985	0.8707	0.8707	0.8451	0.8424	0.8145	0.8117
	20	0.9009	0.9002	0.8508	0.8499	0.8068	0.8060	0.7575	0.7566
	40	0.9015	0.9012	0.8284	0.8280	0.7670	0.7667	0.7027	0.7023
	80	0.9016	0.9015	0.8022	0.8021	0.7246	0.7245	0.6490	0.6488
0.95	10	0.9474	0.9485	0.9248	0.9262	0.9052	0.9067	0.8810	0.8827
	20	0.9479	0.9486	0.9095	0.9105	0.8751	0.8762	0.8306	0.8318
	40	0.9482	0.9485	0.8920	0.8925	0.8382	0.8388	0.7803	0.7810
	80	0.9490	0.9491	0.8697	0.8700	0.8010	0.8013	0.7275	0.7279
0.99	10	0.9863	0.9888	0.9758	0.9796	0.9660	0.9708	0.9536	0.9596
	20	0.9869	0.9881	0.9682	0.9705	0.9488	0.9518	0.9217	0.9257
	40	0.9873	0.9879	0.9588	0.9601	0.9256	0.9275	0.8825	0.8849
	80	0.9878	0.9880	0.9455	0.9462	0.8968	0.8980	0.8387	0.8401

7.5.3 Hypothesis Testing

In this section, Monte Carlo tests of hypotheses are based on the same seven distributions: $N(0, 1)$, $GL(1.0)$, $GL(0.1)$, $GL(0.01)$, $SK(2)$, $SK(3)$, and $SK(25)$. Each simulation is based on $L = 1{,}000{,}000$ bivariate random samples of size $N = 20$ and $N = 80$ for $\rho_{xy} = 0.00$ and $\rho_{xy} = +0.60$ and compared to seven nominal upper-tail probability values of $P = 0.99, 0.90, 0.75, 0.50, 0.25, 0.10$, and 0.01. Two tests of $\rho_{xy} \neq 0.00$ are considered. The first test is based on the Fisher z

Table 7.23 Containment probability values for a bivariate $SK(25)$ distribution with Fisher (F) and Jeyaratnam (J) $1-\alpha$ correlation confidence intervals

		$\rho_{xy} = 0.00$		$\rho_{xy} = +0.40$		$\rho_{xy} = +0.60$		$\rho_{xy} = +0.80$	
$1-\alpha$	N	F	J	F	J	F	J	F	J
0.90	10	0.9009	0.8988	0.9134	0.9114	0.9288	0.9270	0.9485	0.9471
	20	0.9010	0.9003	0.9151	0.9145	0.9322	0.9317	0.9556	0.9552
	40	0.9006	0.9004	0.9159	0.9157	0.9340	0.9338	0.9590	0.9589
	80	0.9005	0.9004	0.9157	0.9156	0.9347	0.9346	0.9605	0.9604
0.95	10	0.9476	0.9487	0.9551	0.9561	0.9648	0.9657	0.9759	0.9765
	20	0.9489	0.9496	0.9577	0.9583	0.9691	0.9696	0.9817	0.9821
	40	0.9496	0.9496	0.9592	0.9595	0.9704	0.9707	0.9844	0.9845
	80	0.9494	0.9496	0.9599	0.9600	0.9716	0.9717	0.9853	0.9854
0.99	10	0.9862	0.9888	0.9889	0.9910	0.9919	0.9935	0.9950	0.9960
	20	0.9881	0.9892	0.9911	0.9921	0.9943	0.9950	0.9973	0.9976
	40	0.9891	0.9897	0.9923	0.9927	0.9951	0.9954	0.9981	0.9982
	80	0.9896	0.9898	0.9925	0.9928	0.9959	0.9960	0.9985	0.9986

transform and uses the standardized test statistic given by

$$T = \frac{z - \mu_z}{\sigma_z} \, ,$$

where

$$z = \tanh^{-1}(r_{xy}) \, , \quad \mu_z = \tanh^{-1}(\rho_{xy}) \, , \quad \text{and} \quad \sigma_z = \frac{1}{\sqrt{N-3}} \, .$$

The second test is based on corrected values proposed by Gayen [17], where

$$z = \tanh^{-1}(r_{xy}) \, ,$$

$$\mu_z = \tanh^{-1}(\rho_{xy}) + \frac{\rho_{xy}}{2(N-1)} \left[1 + \frac{5 - \rho_{xy}^2}{4(N-1)} \right] \, ,$$

and

$$\sigma_z = \left\{ \frac{1}{N-1} \left[1 + \frac{4 - \rho_{xy}^2}{2(N-1)} + \frac{22 - 6\rho_{xy}^2 - 3\rho_{xy}^4}{6(N-1)^2} \right] \right\}^{1/2} \, .$$

The results of the Monte Carlo analyses are summarized in Tables 7.24, 7.25, 7.26, 7.27, 7.28, 7.29, 7.30, which contain simulated upper-tail probability values for the seven distributions with specified nominal probability values of P (0.99, 0.95, 0.75, 0.50, 0.25, 0.10, 0.01), ρ_{xy} (0.00, +0.60), and N (20, 80) for the Fisher

Table 7.24 Upper-tail probability values compared with nominal values (P) for a bivariate $N(0, 1)$ distribution with Fisher (F) and Gayen (G) tests of hypotheses on $\rho_{xy} = 0.00$ and $\rho_{xy} = 0.60$

	$N = 20$				$N = 80$			
	$\rho_{xy} = 0.00$		$\rho_{xy} = 0.60$		$\rho_{xy} = 0.00$		$\rho_{xy} = +0.60$	
P	F	G	F	G	F	G	F	G
0.99	0.9894	0.9893	0.9915	0.9895	0.9898	0.9898	0.9908	0.9899
0.90	0.9016	0.9014	0.9147	0.9022	0.9009	0.9009	0.9065	0.9005
0.75	0.7531	0.7529	0.7754	0.7525	0.7514	0.7514	0.7622	0.7512
0.50	0.5001	0.5001	0.5281	0.4997	0.5008	0.5008	0.5141	0.5006
0.25	0.2464	0.2466	0.2685	0.2471	0.2495	0.2496	0.2601	0.2494
0.10	0.0983	0.0985	0.1098	0.0986	0.0999	0.1000	0.1054	0.0995
0.01	0.0108	0.0108	0.0126	0.0110	0.0102	0.0102	0.0110	0.0101

Table 7.25 Upper-tail probability values compared with nominal values (P) for a bivariate $GL(1.0)$ distribution with Fisher (F) and Gayen (G) tests of hypotheses on $\rho_{xy} = 0.00$ and $\rho_{xy} = +0.60$

	$N = 20$				$N = 80$			
	$\rho_{xy} = 0.00$		$\rho_{xy} = 0.60$		$\rho_{xy} = 0.00$		$\rho_{xy} = 0.60$	
P	F	G	F	G	F	G	F	G
0.99	0.9892	0.9891	0.9878	0.9853	0.9897	0.9897	0.9851	0.9838
0.90	0.9019	0.9016	0.9020	0.8888	0.9011	0.9011	0.8880	0.8817
0.75	0.7539	0.7537	0.7638	0.7419	0.7518	0.7518	0.7451	0.7348
0.50	0.4999	0.4999	0.5324	0.5060	0.5004	0.5004	0.5158	0.5037
0.25	0.2457	0.2460	0.2895	0.2688	0.2495	0.2495	0.2815	0.2715
0.10	0.0981	0.0983	0.1314	0.1197	0.1000	0.1000	0.1290	0.1228
0.01	0.0109	0.0109	0.0195	0.0173	0.0102	0.0102	0.0190	0.0177

Table 7.26 Upper-tail probability values compared with nominal values (P) for a bivariate $GL(0.1)$ distribution with Fisher (F) and Gayen (G) tests of hypotheses on $\rho_{xy} = 0.00$ and $\rho_{xy} = +0.60$

	$N = 20$				$N = 80$			
	$\rho_{xy} = 0.00$		$\rho_{xy} = 0.60$		$\rho_{xy} = 0.00$		$\rho_{xy} = 0.60$	
P	F	G	F	G	F	G	F	G
0.99	0.9918	0.9918	0.9869	0.9841	0.9916	0.9916	0.9819	0.9804
0.90	0.9059	0.9056	0.8954	0.8818	0.9026	0.9026	0.8774	0.8710
0.75	0.7502	0.7499	0.7560	0.7342	0.7484	0.7484	0.7347	0.7247
0.50	0.2436	0.4908	0.5297	0.5045	0.4937	0.4937	0.5144	0.5027
0.25	0.1016	0.2438	0.2982	0.2784	0.2470	0.2470	0.2921	0.2824
0.10	0.0137	0.1018	0.1441	0.1323	0.1016	0.1016	0.1435	0.1373
0.01	0.0000	0.0138	0.0257	0.0231	0.0122	0.0122	0.0265	0.0250

Table 7.27 Upper-tail probability values compared with nominal values (P) for a bivariate $GL(0.01)$ distribution with Fisher (F) and Gayen (G) tests of hypotheses on $\rho_{xy} = 0.00$ and $\rho_{xy} = +0.60$

	$N = 20$				$N = 80$			
	$\rho_{xy} = 0.00$		$\rho_{xy} = 0.60$		$\rho_{xy} = 0.00$		$\rho_{xy} = 0.60$	
P	F	G	F	G	F	G	F	G
0.99	0.9924	0.9923	0.9865	0.9837	0.9920	0.9920	0.9890	0.9792
0.90	0.9060	0.9058	0.8940	0.8803	0.9030	0.9030	0.8740	0.8675
0.75	0.7491	0.7488	0.7544	0.7329	0.7481	0.7481	0.7311	0.7210
0.50	0.4893	0.4893	0.5301	0.5054	0.4921	0.4921	0.5135	0.5018
0.25	0.2429	0.2431	0.3010	0.2810	0.2469	0.2469	0.2947	0.2850
0.10	0.1019	0.1021	0.1476	0.1357	0.1019	0.1019	0.1476	0.1416
0.01	0.0141	0.0142	0.0279	0.0250	0.0128	0.0128	0.0285	0.0268

Table 7.28 Upper-tail probability values compared with nominal values (P) for a bivariate $SK(2)$ distribution with Fisher (F) and Gayen (G) tests of hypotheses on $\rho_{xy} = 0.00$ and $\rho_{xy} = +0.60$

	$N = 20$				$N = 80$			
	$\rho_{xy} = 0.00$		$\rho_{xy} = 0.60$		$\rho_{xy} = 0.00$		$\rho_{xy} = 0.60$	
P	F	G	F	G	F	G	F	G
0.99	0.9842	0.9841	0.9487	0.9423	0.9852	0.9852	0.8480	0.8442
0.90	0.9096	0.9094	0.8159	0.8016	0.9167	0.9167	0.7162	0.7111
0.75	0.7739	0.7737	0.6918	0.6750	0.7838	0.7837	0.6221	0.6165
0.50	0.5001	0.5001	0.5327	0.5163	0.5002	0.5002	0.5121	0.5064
0.25	0.2263	0.2265	0.3797	0.3662	0.2172	0.2172	0.4060	0.4011
0.10	0.0905	0.0907	0.2650	0.2548	0.0834	0.0834	0.3224	0.3182
0.01	0.0159	0.0160	0.1333	0.1284	0.0151	0.0151	0.2099	0.2071

Table 7.29 Upper-tail probability values compared with nominal values (P) for a bivariate $SK(3)$ distribution with Fisher (F) and Gayen (G) tests of hypotheses on $\rho_{xy} = 0.00$ and $\rho_{xy} = +0.60$

	$N = 20$				$N = 80$			
	$\rho_{xy} = 0.00$		$\rho_{xy} = 0.60$		$\rho_{xy} = 0.00$		$\rho_{xy} = 0.60$	
P	F	G	F	G	F	G	F	G
0.99	0.9883	0.9883	0.9766	0.9726	0.9887	0.9887	0.9463	0.9437
0.90	0.9034	0.9032	0.8731	0.8595	0.9031	0.9031	0.8215	0.8152
0.75	0.7559	0.7557	0.7394	0.7192	0.7553	0.7553	0.6941	0.6854
0.50	0.4998	0.4998	0.5348	0.5119	0.4998	0.4998	0.5169	0.5076
0.25	0.2440	0.2442	0.3249	0.3067	0.2450	0.2451	0.3394	0.3315
0.10	0.0967	0.0970	0.1790	0.1672	0.0973	0.0973	0.2107	0.2051
0.01	0.0118	0.0119	0.0506	0.0471	0.0112	0.0112	0.0807	0.0783

Table 7.30 Upper-tail probability values compared with nominal values (P) for a bivariate $SK(25)$ distribution with Fisher (F) and Gayen (G) tests of hypotheses on $\rho_{xy} = 0.00$ and $\rho_{xy} = +0.60$

	$N = 20$				$N = 80$			
	$\rho_{xy} = 0.00$		$\rho_{xy} = 0.60$		$\rho_{xy} = 0.00$		$\rho_{xy} = 0.60$	
P	F	G	F	G	F	G	F	G
0.99	0.9890	0.9889	0.9955	0.9943	0.9899	0.9899	0.9958	0.9953
0.90	0.9014	0.9017	0.9337	0.9217	0.9006	0.9006	0.9292	0.9237
0.75	0.7538	0.7536	0.7928	0.7679	0.7512	0.7512	0.7831	0.7714
0.50	0.5005	0.5005	0.5179	0.4861	0.5004	0.5004	0.5076	0.4924
0.25	0.2463	0.2465	0.2354	0.2133	0.2493	0.2493	0.2295	0.2184
0.10	0.0975	0.0978	0.0830	0.0734	0.0999	0.0999	0.0785	0.0734
0.01	0.0111	0.0112	0.0072	0.0062	0.0103	0.0103	0.0054	0.0049

(F) and Gayen (G) test statistics. Table 7.24 analyzes data obtained from the $N(0, 1)$ distribution; Tables 7.25, 7.26, and 7.27 analyze data obtained from the generalized logistic distribution with $\theta = 1.0$, 0.1, and 0.01, respectively; and Tables 7.28, 7.29, and 7.30 analyze data obtained from the symmetric kappa distribution with $\lambda = 2$, 3, and 25, respectively.

In each table, the Monte Carlo upper-tail probability values for tests of hypotheses based on the Fisher and Gayen approaches were obtained from the same $L = 1{,}000{,}000$ bivariate random samples of size N drawn with replacement from the designated bivariate distribution characterized by the specified population correlation ρ_{xy}. If the Fisher [14] and Gayen [17] techniques are appropriate for the simulated data, the upper-tail probability values should agree with the nominal upper-tail values, P.

Considered as a set, some general statements can be made about the Monte Carlo results contained in Tables 7.24 through 7.30. First, both the Fisher z transform and the Gayen correction provide very satisfactory results for the $N(0, 1)$ distribution analyzed in Table 7.24 with any value of ρ_{xy} and for any of the other distributions analyzed in Tables 7.25 through 7.30 when $\rho_{xy} = 0.00$. Second, in general the Monte Carlo upper-tail probability values obtained with the Gayen correction are better than those obtained with the uncorrected Fisher z transform, especially near $P = 0.50$. Where differences exist, the Fisher z transform is somewhat better than the Gayen correction with $P > 0.75$ and the Gayen correction performs better when $P < 0.75$. Third, discrepancies between the Monte Carlo upper-tail probability values and the nominal probability values are noticeably larger for $N = 80$ than for $N = 20$ and for $\rho_{xy} = 0.60$ than for $\rho_{xy} = 0.00$, especially for the skewed and heavy-tailed distributions, i.e., $GL(0.1)$, $GL(0.01)$, $SK(2)$, and $SK(3)$. Fourth, the Monte Carlo upper-tail probability values in Tables 7.24 through 7.30 are consistently closer to the nominal values for $\rho_{xy} = 0.00$ than for $\rho_{xy} = +0.60$.

To illustrate the difference in results among the seven distributions, consider the first and last values in the last column in each table, i.e., the two Gayen values corresponding to $P = 0.99$ and $P = 0.01$ for $N = 80$ and $\rho_{xy} = +0.60$ in

Tables 7.25 to 7.30, inclusive. If an investigator was to test the null hypothesis H_0: $\rho_{xy} = +0.60$ with a two-tailed test at $\alpha = 0.02$, then given the $N(0, 1)$ distribution analyzed in Table 7.24, the investigator would reject the null hypothesis at a rate of 0.0202 or about 2.02% of the time, i.e., $1.0000 - 0.9899 + 0.0101 = 0.0202$, which is very close to $\alpha = 0.02$. For the light-tailed $GL(1.0)$ or generalized logistic distribution analyzed in Table 7.25, the investigator would reject H_0: $\rho_{xy} = 0.60$ at a rate of 0.0339 or about 3.39% of the time, i.e., $1.0000 - 0.9838 + 0.0177 = 0.0339$, compared with the specified $\alpha = 0.02$. For the skewed $GL(0.1)$ distribution analyzed in Table 7.26, the investigator would reject H_0: $\rho_{xy} = +0.60$ at a rate of 0.0446 or about 4.46% of the time, and for the $GL(0.01)$ distribution analyzed in Table 7.27, which has a more pronounced skewness than $GL(0.1)$, the rejection rate is 0.0476 or about 4.76%, compared to $\alpha = 0.02$. The heavy-tailed distributions, $SK(2)$ and $SK(3)$, analyzed in Tables 7.28 and 7.29, respectively, yield rejection rates of 0.3629 and 0.1346, respectively, which are not the least bit close to $\alpha = 0.02$. Finally, the very light-tailed distribution, $SK(25)$, analyzed in Table 7.30 yields a reversal with a very conservative rejection rate of 0.0096, compared to $\alpha = 0.02$.

7.5.4 Discussion

The Fisher z transform of the sample correlation coefficient, r_{xy}, is widely used in a variety of disciplines for both estimating population ρ_{xy} values and for testing hypothesized values of $\rho_{xy} \neq 0.00$. The transform is presented in most textbooks and is a standard feature of many statistical software packages. The assumptions underlying the use of the Fisher z transform are (1) a simple random sample drawn with replacement from (2) a bivariate normal distribution. It is commonly believed that the Fisher z transform is robust to non-normality. For example, in 1929 Karl Pearson observed:

> [T]he normal bivariate surface can be mutilated and distorted to a remarkable degree without affecting the frequency distribution of r in samples as small as 20 [43, p. 357].

Given correlated non-normal bivariate distributions, these Monte Carlo analyses demonstrate that the Fisher z transform is not at all robust.

In general, while the Fisher z transform and the alternative techniques proposed by Gayen [17] and Jeyaratnam [22] provide accurate results for a bivariate normal distribution with any value of ρ_{xy} and for non-normal bivariate distributions when $\rho_{xy} = 0.0$, serious problems surface with non-normal bivariate distributions when $|\rho_{xy}| > 0.0$. The results for the light-tailed $SK(25)$ distribution are, in general, slightly conservative when $|\rho_{xy}| > 0.0$; cf. Liu, Woodward, and Bonett [28, p. 508]. This is usually not seen as a serious problem in practice as conservative results imply possible failure to reject the null hypothesis and a potential increase in type II error. In comparison, the results for the heavy-tailed distributions, $SK(2)$ and $SK(3)$, and the skewed distributions, $GL(0.1)$ and $GL(0.01)$ are quite liberal when $|\rho_{xy}| > 0.0$.

Also, $GL(1.0)$ is a light-tailed distribution that yields slightly liberal results. Liberal results are much more serious than conservative results, as they imply possible rejection of the null hypothesis and a potential increase in type I error.

Most surprisingly, from a statistical perspective, for the heavy-tailed and skewed distributions, small samples provide better estimates than large samples. Table 7.31 extends the analyses of Tables 7.21, 7.22, 7.23, and 7.24 to larger sample sizes. In Table 7.31 the investigation is limited to Monte Carlo containment probability values obtained from the Fisher z transform for the skewed bivariate distributions based on $GL(0.1)$ and $GL(0.01)$ and for the heavy-tailed bivariate distributions based on $SK(2)$ and $SK(3)$, with $\rho_{xy} = 0.00$ and $\rho_{xy} = 0.60$, and for $N = 10$, 20, 40, 80, 160, 320, and 640. Inspection of Table 7.31 confirms that the trend observed in Tables 7.19 through 7.22 continues with larger sample sizes, producing increasingly smaller containment probability values with increasing N for $|\rho_{xy}| > 0.00$, where $\rho_{xy} = +0.60$ is considered representative of larger ρ_{xy} values.

The impact of large sample sizes is most pronounced in the heavy-tailed bivariate distribution based on $SK(2)$ and the skewed bivariate distribution based on $GL(0.01)$ where, with $\rho_{xy} = +0.60$, the divergence between the containment probability values and the nominal $1 - \alpha$ values for $N = 10$ and $N = 640$ is quite extreme. For example, $SK(2)$ with $1 - \alpha = 0.90$, $\rho_{xy} = +0.60$, and $N = 10$ yields a containment probability value of $P = 0.7487$, whereas $N = 640$ for this case yields a containment probability value of $P = 0.2677$, compared with $1 - \alpha = 0.90$. Obviously, large samples have a greater chance of selecting rare extreme values than small samples. Consequently, the Monte Carlo containment probability values become worse with increasing sample size when heavy-tailed distributions are encountered.

It is clear that the Fisher z transform provides very good results for the bivariate normal distribution and any of the other distributions when $\rho_{xy} = 0.00$. However, if a distribution is not bivariate normal and $\rho_{xy} > 0.00$, then the Fisher z random variable does not follow a normal distribution. Geary [18, p. 241] admonished: "Normality is a myth; there never was, and never will be, a normal distribution." In the absence of bivariate normality and in the presence of correlated heavy-tailed bivariate distributions, such as those contaminated by extreme values, or correlated skewed bivariate distributions, the Fisher z transform and related techniques can yield highly inaccurate results.

Given that normally distributed populations are rarely encountered in actual research situations [18, 33] and that both heavy-tailed symmetrical distributions and heavy-tailed skewed distributions are prevalent in much research, considerable caution should be exercised when using the Fisher z transform or related techniques such as those proposed by Gayen [17] and Jeyaratnam [22], as these methods clearly are not robust to deviations from normality when $|\rho_{xy}| \neq 0.0$. In general, there is no easy answer to this problem. However, a researcher cannot simply ignore a problem just because it is annoying. Unfortunately, given a non-normal population with $\rho_{xy} \neq 0.0$, there appear to be no published alternative tests of significance nor viable options for the construction of confidence intervals.

Table 7.31 Containment probability values for the bivariate $GL(0.1)$, $GL(0.01)$, $SK(2)$, and $SK(3)$ distributions with Fisher (F) $1-\alpha$ correlation confidence intervals

		Distribution							
		$GL(0.1)$		$GL(0.01)$		$SK(2)$		$SK(3)$	
$1-\alpha$	N	$\rho_{xy} = 0.00$	$\rho_{xy} = 0.60$	$\rho_{xy} = 0.00$	$\rho_{xy} = 0.60$	$\rho_{xy} = 0.00$	$\rho_{xy} = +0.60$	$\rho_{xy} = 0.00$	$\rho_{xy} = +0.60$
0.90	10	0.9016	0.8729	0.9019	0.8693	0.8961	0.7487	0.9007	0.8451
	20	0.9013	0.8593	0.9015	0.8545	0.9002	0.6650	0.9009	0.8068
	40	0.9010	0.8510	0.9012	0.8454	0.9050	0.5755	0.9015	0.7670
	80	0.9006	0.8459	0.9002	0.8394	0.9097	0.4884	0.9016	0.7246
	160	0.9004	0.8431	0.9004	0.8366	0.9138	0.4060	0.9021	0.6822
	320	0.9003	0.8405	0.9003	0.8338	0.9173	0.3314	0.9025	0.6369
	640	0.9002	0.8400	0.9001	0.8332	0.9204	0.2677	0.9016	0.5934
0.95	10	0.9486	0.9281	0.9485	0.9255	0.9403	0.8217	0.9474	0.9052
	20	0.9495	0.9197	0.9496	0.9160	0.9415	0.7457	0.9479	0.8751
	40	0.9495	0.9136	0.9495	0.9092	0.9436	0.6551	0.9482	0.8382
	80	0.9498	0.9100	0.9500	0.9055	0.9461	0.5634	0.9490	0.8010
	160	0.9504	0.9075	0.9503	0.9025	0.9490	0.4714	0.9495	0.7590
	320	0.9500	0.9063	0.9500	0.9011	0.9514	0.3889	0.9497	0.7164
	640	0.9498	0.9053	0.9499	0.9001	0.9535	0.3147	0.9500	0.6714
0.99	10	0.9871	0.9793	0.9869	0.9782	0.9797	0.9152	0.9863	0.9660
	20	0.9882	0.9770	0.9881	0.9752	0.9789	0.8577	0.9869	0.9488
	40	0.9890	0.9752	0.9889	0.9732	0.9788	0.7780	0.9873	0.9256
	80	0.9895	0.9737	0.9897	0.9712	0.9794	0.6835	0.9878	0.8968
	160	0.9896	0.9726	0.9896	0.9702	0.9802	0.5854	0.9877	0.8639
	320	0.9899	0.9721	0.9899	0.9697	0.9811	0.4901	0.9883	0.8272
	640	0.9900	0.9721	0.9899	0.9696	0.9817	0.4020	0.9885	0.7877

Finally, to paraphrase a line from Thompson regarding the use of tiltmeters in volcanology [53, p. 258],

1. Do not use the Fisher z transformation.
2. If you do use it, don't believe it.
3. If you do believe it, don't publish it.
4. If you do publish it, don't be the first author.

7.6 Point-Biserial Linear Correlation

The point-biserial correlation coefficient measures the association between a dichotomous variable and an interval-level variable. Applications of the point-biserial correlation abound in fields such as education and educational psychology. The point-biserial correlation may be thought of simply as the Pearson product-moment correlation between an interval-level variable and a variable with two disjoint, unordered categories.

7.6.1 Example

To illustrate the point-biserial correlation coefficient, consider the dichotomous data listed in Table 7.32 for $N = 13$ subjects where variable x is a dichotomous variable coded $(0, 1)$ and variable y is an interval-level variable. The point-biserial correlation is usually computed as

$$r_{pb} = \frac{\bar{y}_1 - \bar{y}_0}{s_y} \sqrt{\frac{n_0 n_1}{N(N-1)}} \, ,$$

Table 7.32 Example bivariate data for point-biserial correlation on $N = 13$ subjects

Subject	x	y
1	0	19
2	1	17
3	0	18
4	0	18
5	1	26
6	1	28
7	0	20
8	1	19
9	0	22
10	1	23
11	1	26
12	0	25
13	1	30

where n_0 and n_1 denote the number of y values coded 0 and 1, respectively, $N = n_0 + n_1$, $\bar{y}_0$ and $\bar{y}_1$ denote the means of the y values coded 0 and 1, respectively, and s_y is the sample standard deviation of the y values given by

$$s_y = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2} \, .$$

For the data listed in Table 7.32, $n_0 = 6$, $n_1 = 7$, $\bar{y}_0 = 20.3333$, $\bar{y}_1 = 24.1429$, $s_y = 4.2728$, and the point-biserial correlation is

$$r_{pb} = \frac{\bar{y}_1 - \bar{y}_0}{s_y}\sqrt{\frac{n_0 n_1}{N(N-1)}} = \frac{24.1429 - 20.3333}{4.2728}\sqrt{\frac{(6)(7)}{13(13-1)}} = +0.4626 \, .$$

However, r_{pb} can also be calculated simply as the Pearson product-moment correlation (r_{xy}) between dichotomous variable x and interval variable y. For the data listed in Table 7.32, $N = 13$,

$$\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} x_i^2 = 7 \, , \quad \sum_{i=1}^{N} y_i = 291 \, , \quad \sum_{i=1}^{N} y_i^2 = 6{,}733 \, , \quad \sum_{i=1}^{N} x_i y_i = 169 \, ,$$

and

$$r_{xy} = \frac{N\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{\sqrt{\left[N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2\right]\left[N\sum_{i=1}^{N} y_i^2 - \left(\sum_{i=1}^{N} y_i\right)^2\right]}}$$

$$= \frac{(13)(169) - (7)(291)}{\sqrt{\left[(13)(7) - 7^2\right]\left[(13)(6{,}733) - 291^2\right]}} = +0.4626 \, .$$

Approaching the calculation of the probability value from a product-moment perspective, there are

$$M = N! = 13! = 6{,}227{,}020{,}800$$

possible, equally-likely arrangements in the reference set of all permutations of the observed bivariate data, making an exact permutation analysis impractical. Let r_o denote the observed value of r_{pb}. Then, based on $L = 1{,}000{,}000$ random arrangements of the observed data under the null hypothesis, there are 121,667 $|r_{pb}|$

values equal to or greater than $|r_o| = 0.4626$, yielding a Monte Carlo resampling two-sided probability value of $P = 121{,}667/1{,}000{,}000 = 0.121667$.

In general, $L = 1{,}000{,}000$ ensures three decimal places of accuracy. However, it requires an increase of two orders of magnitude, i.e., $L = 100{,}000{,}000$, to ensure four decimal places of accuracy [23]. Based on $L = 100{,}000{,}000$ random arrangements of the observed bivariate data, the two-sided Monte Carlo resampling probability value of $r_{pb} = +0.4626$ to six decimal places is $P = 12{,}121{,}600/100{,}000{,}000 = 0.121216$.

However, because variable x is composed of only two categories, an alternative procedure exists for establishing the probability value of r_{pb}. The relationships between r_{pb} and Student's two-sample t test are

$$r_{pb} = \sqrt{\frac{t^2}{t^2 + N - 2}} \quad \text{and} \quad t = \frac{r_{pb}\sqrt{N-2}}{\sqrt{1 - r_{pb}^2}} .$$

Thus, the probability value for a specified point-biserial correlation coefficient can be calculated much more efficiently as the probability value of a two-sample t test with $N - 2$ degrees of freedom. Consider the data in Table 7.32 rearranged into two groups coded 0 and 1 as in Table 7.33.

For the observed data listed in Table 7.33, Student's t test statistic is

$$t = \frac{r_{pb}\sqrt{N-2}}{\sqrt{1 - r_{pb}^2}} = \frac{+0.4626\sqrt{13-2}}{\sqrt{1 - (+0.4626)^2}} = +1.7307 .$$

For the data listed in Table 7.33, there are only

$$M = \frac{N!}{n_0!\, n_1!} = \frac{13!}{6!\, 7!} = 1{,}716$$

possible, equally-likely arrangements in the reference set of all permutations of the observed scores, compared with

$$M = N! = 13! = 6{,}227{,}020{,}800$$

Table 7.33 Example data on $N = 13$ subjects for Student's t test

0	1
19	17
18	26
18	28
20	19
22	23
25	26
	30

in the initial set, making an exact permutation analysis possible. If all arrangements of the $N = 13$ observed scores occur with equal chance, the exact two-sided probability value of $t = +1.7307$ to six places computed on the $M = 1{,}716$ possible arrangements of the observed data with $n_0 = 6$ and $n_1 = 7$ preserved for each arrangement is $208/1{,}716 = 0.121212$.

The Monte Carlo resampling probability value of $P = 0.121667$ based on $L = 1{,}000{,}000$ and the Monte Carlo resampling probability value of $P = 0.121216$ based on $L = 100{,}000{,}000$ both compare favorably with the exact probability value of $P = 0.121212$. For comparison, the two-sided probability value of $t = +1.7303$ based on Student's t distribution with $N - 2 = 13 - 2 = 11$ degrees of freedom is $P = 0.111421$.

7.6.2 Problems with the Point-Biserial Coefficient

Whenever a dichotomous variable is correlated with an interval-level variable, as in point-biserial correlation, there are potential problems with proper norming between ± 1. In brief, it is not possible to obtain a perfect correlation, positive or negative, between a dichotomous variable and a continuous variable [42, p. 145]. The reason is simply that it is not possible for a dichotomous variable and a continuous variable to have the same shape, as illustrated in Fig. 7.8 where a dichotomous variable (x) is correlated with a continuous variable (y) that is comprised of a uniform distribution, i.e, $y = 1, 2, \ldots, 10$. In order to achieve a perfect correlation of $r_{pb} = +1.00$, it

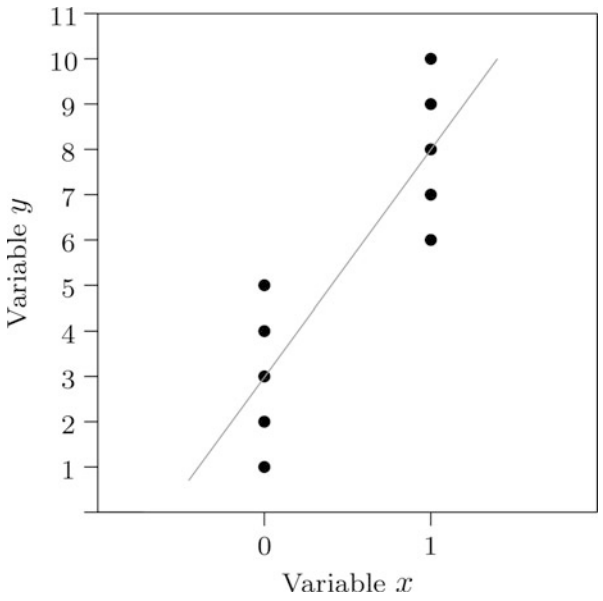

Fig. 7.8 Scatterplot of a uniform distribution of y values with the regression line overlaid

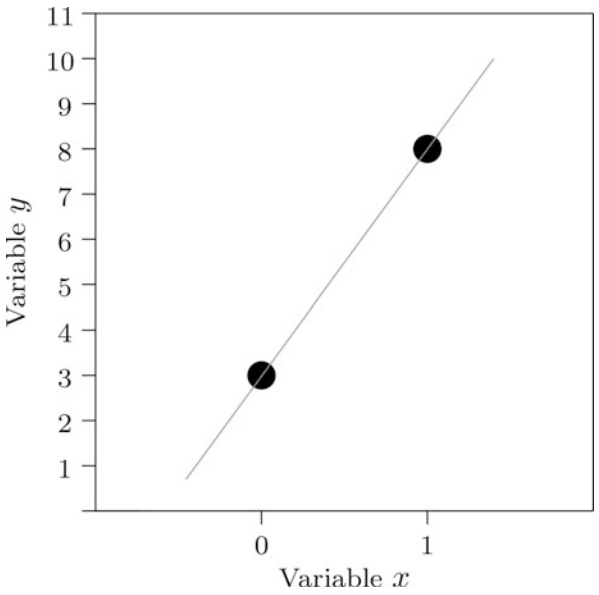

Fig. 7.9 Scatterplot of clusters of y values located at $x = 0$ and $x = 1$ with the regression line overlaid

would be necessary for all the scores at the two points of variable x ($x = 0$ and $x = 1$) to fall exactly on two points on variable y, as depicted in Fig. 7.9 where the larger black circles represent a cluster of points at $x = 0$ and $x = 1$. Since variable y is assumed to be continuous, this is not possible. Consequently, values of variable y at either of the two points on variable x (the dichotomous variable) must correspond to a range of points on variable y (the continuous variable).

As Jum Nunnally showed in 1978, the maximum value of r_{pb} between a dichotomous variable and a normally distributed variable is approximately $r_{pb} = \pm 0.80$, which occurs only when $p = n_0/N = 0.50$ [42]. As p deviates from 0.50 in either direction, the maximum value of r_{pb} is further reduced. Consequently, when $p = 0.25$ or $p = 0.75$, the maximum value of r_{pb} is approximately $r_{pb} = \pm 0.75$, and when $p = 0.90$ or $p = 0.10$, the maximum value of r_{pb} is only approximately $r_{pb} = \pm 0.58$.[5]

The problem can be illustrated with a small empirical example. Table 7.34 contains 10 scores $(1, 2, \ldots, 10)$ with frequencies corresponding to an expanded binomial distribution, which approximates a normal distribution with $N = 512$. For

[5]The problem is not confined to r_{pb}. In general, the problem is called the base-rate problem or the marginal-dependent problem. See two excellent discussions of the problem by Goodman [19] and McGrath and Meyer [31].

Table 7.34 Example binomial distribution on $N = 512$ subjects with $p = 0.50$

x	y	f	fy	y^2	fy^2
0	1	1	1	1	1
0	2	9	18	4	36
0	3	36	108	9	324
0	4	84	336	16	1,344
0	5	126	630	25	3,150
1	6	126	756	36	4,536
1	7	84	588	49	4,116
1	8	36	288	64	2,304
1	9	9	81	81	729
1	10	1	10	100	100
Sum		512	2,816		16,640

the binomial data listed in Table 7.34 with $p = 0.50$,

$$\bar{y}_0 = \left(\sum_{i=1}^{n_0} f_i\right)^{-1} \sum_{i=1}^{n_0} f_i y_i = \frac{1 + 18 + 108 + 336 + 630}{1 + 9 + 36 + 84 + 126} = 4.2695 ,$$

$$\bar{y}_1 = \left(\sum_{i=1}^{n_1} f_i\right)^{-1} \sum_{i=1}^{n_1} f_i y_i = \frac{756 + 588 + 288 + 81 + 10}{126 + 84 + 36 + 9 + 1} = 6.7305 ,$$

$$s_y = \sqrt{\frac{\displaystyle\sum_{i=1}^{N} fy^2 - \frac{\left(\displaystyle\sum_{i=1}^{N} fy\right)^2}{N}}{N - 1}} = \sqrt{\frac{16{,}640 - \dfrac{(2{,}816)^2}{512}}{512 - 1}} = 1.5015 ,$$

and

$$r_{pb} = \frac{\bar{y}_1 - \bar{y}_0}{s_y} \sqrt{\frac{n_0 n_1}{N(N-1)}} = \frac{6.7305 - 4.2695}{1.5015} \sqrt{\frac{(256)(256)}{512(512-1)}}$$

$$= +0.8203 ,$$

which approximates Nunnally's estimate of $r_{pb} = +0.80$.

Table 7.35 illustrates a binomial distribution with $N = 512$ and $p \simeq 0.25$, i.e.,

$$p = \frac{1}{N} \sum_{i=1}^{n_0} f_i = \frac{1 + 9 + 36 + 84}{512} = 0.2539 .$$

Table 7.35 Example binomial distribution on $N = 512$ subjects with $p \simeq 0.25$

x	y	f	fy	y^2	fy^2
0	1	1	1	1	1
0	2	9	18	4	36
0	3	36	108	9	324
0	4	84	336	16	1,344
1	5	126	630	25	3,150
1	6	126	756	36	4,536
1	7	84	588	49	4,116
1	8	36	288	64	2,304
1	9	9	81	81	729
1	10	1	10	100	100
Sum		512	2,816		16,640

For the binomial data in Table 7.35 with $p \simeq 0.25$,

$$\bar{y}_0 = \left(\sum_{i=1}^{n_0} f_i \right)^{-1} \sum_{i=1}^{n_0} f_i y_i = \frac{1 + 18 + 108 + 336}{1 + 9 + 36 + 84} = 3.5615 ,$$

$$\bar{y}_1 = \left(\sum_{i=1}^{n_1} f_i \right)^{-1} \sum_{i=1}^{n_1} f_i y_i = \frac{630 + 756 + 588 + 288 + 81 + 10}{126 + 126 + 84 + 36 + 9 + 1} = 6.1597 ,$$

the standard deviation of the y values is unchanged at $s_y = 1.5015$ and

$$r_{pb} = \frac{\bar{y}_1 - \bar{y}_0}{s_y} \sqrt{\frac{n_0 n_1}{N(N-1)}} = \frac{6.1597 - 3.5615}{1.5015} \sqrt{\frac{(130)(382)}{512(512-1)}}$$

$$= +0.7539 ,$$

which approximates Nunnally's estimate of $r_{pb} = +0.75$.

While it is not convenient to take exactly 10% of $N = 512$ cases, as arranged in Table 7.34, it is possible to take 9% of $N = 512$ cases. Thus,

$$p = \frac{1}{N} \sum_{i=1}^{n_0} = \frac{1 + 9 + 36}{512} = \frac{46}{512} = 0.0898.$$

Table 7.36 Example binomial distribution on $N = 512$ subjects with $p = 0.09$

x	y	f	fy	y^2	fy^2
0	1	1	1	1	1
0	2	9	18	4	36
0	3	36	108	9	324
1	4	84	336	16	1,344
1	5	126	630	25	3,150
1	6	126	756	36	4,536
1	7	84	588	49	4,116
1	8	36	288	64	2,304
1	9	9	81	81	729
1	10	1	10	100	100
Sum		512	2,816		16,640

Table 7.36 illustrates a binomial distribution with $N = 512$ and $p = 0.09$. For the binomial data listed in Table 7.36 with $p \simeq 0.10$,

$$\bar{y}_0 = \left(\sum_{1=1}^{n_0} f_i \right)^{-1} \sum_{i=1}^{n_0} f_i y_i = \frac{1 + 18 + 108}{1 + 9 + 36} = 2.7609 ,$$

$$\bar{y}_1 = \left(\sum_{i=1}^{n_1} f_i \right)^{-1} \sum_{i=1}^{n_1} f_i y_i = \frac{336 + 630 + 756 + 588 + 288 + 81 + 10}{84 + 126 + 126 + 84 + 36 + 9 + 1}$$

$$= 5.7704 ,$$

the standard deviation of the y values is unchanged at $s_y = 1.5015$ and

$$r_{pb} = \frac{\bar{y}_1 - \bar{y}_0}{s_y} \sqrt{\frac{n_0 n_1}{N(N-1)}} = \frac{5.7704 - 2.7609}{1.5015} \sqrt{\frac{(46)(466)}{512(512-1)}}$$

$$= +0.5737 ,$$

which approximates Nunnally's estimate of $r_{pb} = +0.58$.

7.7 Biserial Linear Correlation

Point-biserial correlation measures the degree of association between an interval-level variable and a dichotomous variable that is a true dichotomy, such as right and wrong, true and false, or left and right. On the other hand, biserial correlation measures the degree of association between an interval-level variable and a dichotomous variable that has been created from a variable that is assumed to be continuous

and normally distributed, such as grades that have been dichotomized into "pass" and "fail" or weight that has been classified into "normal" and "obese."[6] Biserial correlation has long been difficult to compute, requiring the ordinate of a unit-normal distribution. Some approximating methods have been suggested to simplify computation [16], but these are unnecessary with permutation methods.

Let x represent the dichotomous variable and y represent the continuous interval-level variable, then the biserial correlation coefficient is given by

$$r_b = \frac{(\bar{y}_1 - \bar{y}_0)pq}{uS_y} ,$$

where p and $q = 1 - p$ denote the proportions of all y values coded 0 and 1, respectively, $\bar{y}_0$ and $\bar{y}_1$ denote the arithmetic means of the y values coded 0 and 1, respectively, S_y is the standard deviation of the y values given by[7]

$$S_y = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2} ,$$

and u is the ordinate of the unit normal curve at the point of division between the p and q proportions under the curve given by

$$u = \frac{\exp(-z^2/2)}{\sqrt{2\pi}} .$$

Written in raw terms without the p and q proportions,

$$r_b = \frac{(\bar{y}_0 - \bar{y}_1)n_0 n_1}{N^2 u S_y} ,$$

where n_0 and n_1 denote the number of y values coded 0 and 1, respectively, and $N = n_0 + n_1$. The biserial correlation may also be written in terms of the point-biserial correlation coefficient,

$$r_b = \frac{r_{pb}\sqrt{pq}}{u} = \frac{r_{pb}\sqrt{n_0 n_1}}{Nu} ,$$

[6] For many years height has been considered as normally distributed, but recent research indicates that this is not necessarily the case [30, pp. 205–207].

[7] Note that the sum of squared deviation is divided by N, not $N - 1$ and the symbol for the standard deviation is S_y with an uppercase letter S to distinguish it from the usual sample standard deviation denoted by s_y.

where the point-biserial correlation coefficient is given by

$$r_{pb} = \frac{(\bar{y}_1 - \bar{y}_0)\sqrt{pq}}{S_y} .$$

7.7.1 Example

To illustrate the calculation of the biserial correlation coefficient, consider the set of data given in Table 7.37 where $N = 15$ subjects are scored on interval-level variable y and are classified into types on dichotomous variable x. For the data listed in Table 7.37, $n_0 = 6$, $n_1 = 9$, $p = 6/15 = 0.40$, $q = 9/15 = 0.60$,

$$\bar{y}_0 = \frac{1}{n_0}\sum_{i=1}^{n_0} y_i = \frac{12 + 15 + 11 + 18 + 13 + 11}{6} = 13.3333 ,$$

$$\bar{y}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i = \frac{10 + 33 + 19 + 21 + 29 + 12 + 19 + 23 + 16}{9} = 20.2222 ,$$

$$S_y = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2} = \sqrt{\frac{649.7333}{15}} = 6.5815 ,$$

the standard score that defines the lower $p = 0.40$ of the unit-normal distribution is $z = -0.2533$,

$$u = \frac{\exp(-z^2/2)}{\sqrt{2\pi}} = \frac{\exp[-(-0.2533)^2/2]}{\sqrt{(2)(3.1416)}} = 0.3863 ,$$

and

$$r_b = \frac{(\bar{y}_1 - \bar{y}_0)pq}{uS_y} = \frac{(20.2222 - 13.3333)(0.40)(0.60)}{(0.3863)(6.5815)} = +0.6503 .$$

For the data listed in Table 7.37, the point-biserial correlation coefficient is

$$r_{pb} = \frac{(\bar{y}_1 - \bar{y}_0)\sqrt{pq}}{S_y} = \frac{(20.2222 - 13.3333)\sqrt{(0.40)(0.60)}}{6.5815} = +0.5128 ,$$

and in terms of the point-biserial correlation coefficient, the biserial correlation coefficient is

$$r_b = \frac{r_{pb}\sqrt{pq}}{u} = \frac{+0.5128\sqrt{(0.40)(0.60)}}{0.3863} = +0.6503 .$$

Table 7.37 Example biserial correlation data on $N = 15$ subjects

Subject	x	y
1	0	12
2	0	15
3	0	11
4	0	18
5	0	13
6	0	11
7	1	10
8	1	33
9	1	19
10	1	21
11	1	29
12	1	12
13	1	19
14	1	23
15	1	16

For the $N = 15$ scores listed in Table 7.37, there are only

$$M = \frac{N!}{n_0!\, n_1!} = \frac{15!}{6!\, 9!} = 5{,}005$$

possible, equally-likely arrangements in the reference set of all permutations of the observed scores, making an exact permutation analysis easily accomplished. Note that in the formula for the biserial correlation coefficient,

$$r_b = \frac{\bar{y}_1 - \bar{y}_0 pq}{uS_y}$$

p, q, u, and S_y are invariant under permutation. Therefore, the permutation distribution can efficiently be based entirely on $\bar{y}_1 - \bar{y}_0$. If all $M = 5{,}005$ arrangements of the $N = 15$ observed values occur with equal chance, the exact two-sided probability value of $|r_b| = +0.6503$ computed on the $M = 5{,}005$ possible arrangements of the observed data with $n_0 = 6$ and $n_1 = 9$ preserved for each arrangement is $P = 263/5{,}005 = 0.0525$.

7.8 Intraclass Correlation

There exists an extensive, and controversial, literature on the intraclass correlation coefficient and its uses. The standard reference is by E.A. Haggard, *Intraclass Correlation and the Analysis of Variance* [20], although it has been heavily criticized for both its exposition and its statistical accuracy [51]. See also discussions by

Bartko [3, 2, 4], Kraemer [27], Kraemer and Thiemann [26, pp. 32–34, 54–56], Shrout and Fleiss [50], von Eye and Mun [54, pp. 116-122], and Winer [56, pp. 289–296].

The intraclass correlation coefficient is most often used for measuring the level of agreement among judges. The coefficient represents concordance, where $+1$ indicates perfect agreement and 0 indicates no agreement. While the maximum value of the intraclass correlation coefficient is $+1$, the minimum is given by $-1/(k-1)$, where k is the number of judges. Thus, for $k = 2$ judges the lower limit is -1, but for $k = 3$ judges the lower limit is $-1/2$, for $k = 4$ judges the lower limit is $-1/3$, for $k = 5$ judges the lower limit is $-1/4$, and so on, approaching zero as the number of judges increases. A number of authors recommend that when the intraclass correlation coefficient is negative, it should be interpreted as zero [4, 20, p. 71], but this seems intuitively wrong.

In many ways the intraclass correlation coefficient is a special form of the Pearson product-moment (interclass) correlation coefficient. Consider the small set of data given in Table 7.38 with $N = 5$ subjects and measurements on Height (x) and Weight (y). For the bivariate data given in Table 7.38 with $N = 5$ subjects,

$$\sum_{i=1}^{N} x_i = 15\,, \quad \sum_{i=1}^{N} x_i^2 = 55\,, \quad \sum_{i=1}^{N} y_i = 25\,, \quad \sum_{i=1}^{N} y_i^2 = 135\,, \quad \sum_{i=1}^{N} x_i y_i = 83\,,$$

and the Pearson product-moment correlation coefficient is $r_{xy} = +0.80$.

Now consider $N = 5$ sets of twins and let the variable under consideration be Weight, as in Table 7.39. The question is, which of the two variables labeled Weight is to be considered variable x and which is to be considered variable y? The problem can be solved by the intraclass correlation coefficient using double entries. The intraclass correlation between N pairs of observations on two variables, x and y, is by definition the ordinary Pearson product-moment (interclass) correlation between $2N$ pairs of observations, the first N of which are the original observations, and the

Table 7.38 Example bivariate correlation data on $N = 5$ subjects

Subject	Height (x)	Weight (y)
A	1	4
B	2	3
C	3	5
D	4	7
E	5	6

Table 7.39 Example bivariate correlation data on $N = 5$ twins

Twins	Weight	Weight
A	1	4
B	2	3
C	3	5
D	4	7
E	5	6

second N the original observations with variable x replacing variable y and vice versa [15, Sect. 38]. Table 7.40 illustrates the arrangement. For the bivariate data given in Table 7.40 with $2N = 10$ subjects,

$$\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i = 40\,, \quad \sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} y_i^2 = 190\,, \quad \sum_{i=1}^{N} x_i y_i = 166\,,$$

and the intraclass correlation coefficient is $r_{\mathrm{I}} = +0.20$. Note that certain computational simplifications follow from the reversal of the variables, mainly because the reversals make the marginal distributions for the new variables the same and, therefore, the means and variances of the new variables are also the same [46, p. 20].

For cases with $k > 2$, the construction of a table suitable for calculating the intraclass correlation coefficient is more laborious. For example, given $k = 3$ judges, designate the three values for each subject as x_1, x_2, and x_3. The three values are entered into the table as six observations, each being one of the six permutations of two values that can be made from the original three values. That is, the values of the three values x_1, x_2, and x_3 for each subject are entered into a bivariate correlation table with coordinates (x_1, x_2), (x_1, x_3), (x_2, x_3), (x_2, x_1), (x_3, x_1), and (x_3, x_2), and the Pearson product-moment correlation coefficient is computed for the resulting table, yielding the intraclass correlation coefficient.

To illustrate, consider the small data set given in Table 7.41 with $N = 3$ subjects and $k = 3$ judges. The permutations of the observations in Table 7.41 are listed in the correlation matrix given in Table 7.42. For the bivariate data listed in Table 7.42

Table 7.40 Example bivariate correlation data on $2N = 10$ twins

Twins	Weight (x)	Weight (y)
A	1	4
B	2	3
C	3	5
D	4	7
E	5	6
A′	4	1
B′	3	2
C′	5	3
D′	7	4
E′	6	5

Table 7.41 Example correlation data with $k = 3$ judges and $N = 3$ subjects

Subject	x_1	x_2	x_3
A	1	2	3
B	6	4	5
C	8	9	7

Table 7.42 Bivariate permutation matrix for $k = 3$ judges and $N = 3$ subjects

Ss	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
x	1	1	2	6	6	4	8	8	9	3	3	2	5	5	4	7	7	9
y	2	3	3	4	5	5	9	7	7	1	2	1	4	6	6	9	8	8

Table 7.43 Example data for Case 1, Form 1, with $N = 6$ subjects (S) and $k = 4$ judges (A)

	Judge (A)			
Subject (S)	1	2	3	4
1	9	2	5	8
2	6	1	3	2
3	8	4	6	8
4	7	1	2	6
5	10	5	6	9
6	6	2	4	7

with $N = 18$ subjects,

$$\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i = 90\,, \quad \sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} y_i^2 = 570\,, \quad \sum_{i=1}^{N} x_i\, y_i = 552\,,$$

and the intraclass correlation coefficient obtained via the Pearson product-moment correlation coefficient is $r_{\mathrm{I}} = r_{xy} = +0.85$.

Because of the complexity of double entries with $k > 2$, the intraclass correlation coefficient is usually formulated as an analysis of variance with variable A a random variable. There are actually three different intraclass correlation coefficients, and two forms of each [32, 50, 57]. The three types and two forms are designated as:

$$\mathrm{ICC}(1,1) \text{ and } \mathrm{ICC}(1,k),$$

$$\mathrm{ICC}(2,1) \text{ and } \mathrm{ICC}(2,k),$$

$$\mathrm{ICC}(3,1) \text{ and } \mathrm{ICC}(3,k).$$

Case 1, Form 1: ICC(1, 1) For Case 1, Form 1, there exists a pool of judges. For each subject, a researcher randomly samples k judges from the pool to evaluate each subject. The k judges who rate Subject 1 are not necessarily the same judges who rate Subject 2. To illustrate Case 1, Form 1, Table 7.43 lists example data for $k = 4$ judges (A) and $N = 6$ subjects (S).

Now consider the data given in Table 7.43 as a one-way randomized-block analysis of variance, given in Table 7.44. For the summary data given in Table 7.44, let a indicate the number of levels of Factor A, then the sum-of-squares Total is

$$SS_{\mathrm{Total}} = \sum_{i=1}^{N} x_i^2 - \frac{\left(\sum_{i=1}^{N} x_i\right)^2}{Na} = 841 - \frac{(127)^2}{(6)(4)} = 168.9583\,,$$

Table 7.44 Example data for Case 1, Form 1, prepared for an analysis of variance with $N = 6$ subjects (S) and $k = 4$ judges (A)

	Judge (A)				
Subject (S)	1	2	3	4	T_S
1	9	2	5	8	24
2	6	1	3	2	12
3	8	4	6	8	26
4	7	1	2	6	16
5	10	5	6	9	30
6	6	2	4	7	19
N	6	6	6	6	24
T_A	46	15	26	40	127
Σx^2	366	51	126	298	841

the sum-of-squares Between Subjects (BS) is

$$SS_{\text{BS}} = \frac{\sum_{i=1}^{N} T_{S_i}^2}{a} - \frac{\left(\sum_{i=1}^{N} x_i\right)^2}{Na}$$
$$= \frac{(24)^2 + (12)^2 + \cdots + (19)^2}{4} - \frac{(127)^2}{(6)(4)} = 56.2083 ,$$

the sum-of-squares for Factor A is

$$SS_{\text{A}} = \frac{\sum_{j=1}^{a} T_{A_j}^2}{N} - \frac{\left(\sum_{i=1}^{N} x_i\right)^2}{Na}$$
$$= \frac{(46)^2 + (15)^2 + (26)^2 + (40)^2}{6} - \frac{(127)^2}{(6)(4)} = 97.4583 ,$$

the sum-of-squares Within Subjects (WS) is

$$SS_{\text{WS}} = SS_{\text{Total}} - SS_{\text{BS}} = 168.9583 - 56.2083 = 112.7500 ,$$

and the sum-of-squares Error is

$$SS_{\text{Error}} = SS_{\text{A}\times\text{S}} = SS_{\text{WS}} - SS_{\text{A}} = 112.7500 - 97.4583 = 15.2917 .$$

Table 7.45 Analysis of variance source table for the data given in Table 7.44 with $k = 4$ judges and $N = 6$ subjects

Source	*SS*	*df*	*MS*	*F*
Between *S*	56.2083	5	11.2417	
Within *S*	112.7500	18	6.2639	
Factor *A*	97.4583	3	32.4861	31.87
Error ($A \times S$)	15.2917	15	1.0194	
Total	168.9583	23		

The analysis of variance source table is given in Table 7.45. For Case 1, Form 1, the intraclass correlation coefficient is given by

$$\text{ICC}(1,1) = \frac{MS_{\text{BS}} - MS_{\text{WS}}}{MS_{\text{BS}} + (a-1)MS_{\text{WS}}} = \frac{11.2417 - 6.2639}{11.2417 + (4-1)(6.2639)} = +0.1659\,.$$

Case 1, Form *k*: ICC(1,*k*) If each judge is replaced with a group of k judges, such as a team of clinicians, and the score is the average score of the k judges, then for Case 1, Form k, the intraclass correlation coefficient is

$$\text{ICC}(1,k) = \frac{MS_{\text{BS}} - MS_{\text{WS}}}{MS_{\text{BS}}} = \frac{11.2417 - 6.2639}{11.2417} = +0.4428\,.$$

Case 2, Form 1: ICC(2, 1) If the same set of k judges rate each subject and the k judges are considered a random sample from a population of potential judges, then the intraclass correlation coefficient is designated ICC(2, 1). Because this is the most common case/form, it is usually designated simply as r_{I} in the literature.

$$\text{ICC}(2,1) = \frac{MS_{\text{BS}} - MS_{\text{A}\times\text{S}}}{MS_{\text{BS}} + (a-1)MS_{\text{A}\times\text{S}} + \dfrac{a(MS_{\text{A}} - MS_{\text{A}\times\text{S}})}{N}} = \frac{11.2417 - 1.0194}{11.2417 + (4-1)(1.0194) + \dfrac{(4)(32.4861 - 1.0194)}{6}} = +0.2898\,.$$

Case 2, Form *k*: ICC(2, *k*) If each judge is replaced with a team of k judges, and the score is the average score of the k judges, then for Case 2, Form k, the intraclass correlation coefficient is

$$\text{ICC}(2,k) = \frac{MS_{\text{BS}} - MS_{\text{A}\times\text{S}}}{MS_{\text{BS}} + \dfrac{MS_{\text{A}} - MS_{\text{A}\times\text{S}}}{N}} = \frac{11.2417 - 1.0194}{11.2417 + \dfrac{32.4861 - 1.0194}{6}} = +0.6200\,.$$

Case 3, Form 1: ICC(3, 1) Case 3, Form 1 is the same as Case 2, Form 1, except that the raters are considered as fixed, not random. For Case 3, Form 1, the intraclass correlation coefficient is

$$\text{ICC}(3,1) = \frac{MS_{\text{BS}} - MS_{\text{A}\times\text{S}}}{MS_{\text{BS}} + (a-1)MS_{\text{A}\times\text{S}}} = \frac{11.2417 - 1.0194}{11.2417 + (4-1)(1.0194)} = +0.7148\,.$$

Case 3, Form k: ICC(3, *k*) If each judge is replaced with a team of k judges and the teams are considered as fixed, not random, the intraclass correlation coefficient is

$$\text{ICC}(3,k) = \frac{MS_{\text{BS}} - MS_{\text{A}\times\text{S}}}{MS_{\text{BS}}} = \frac{11.2417 - 1.0194}{11.2417} = +0.9093\,.$$

7.8.1 Example

For another example of the intraclass correlation coefficient, consider Case 2, Form 1, the most common in the literature, with k judges randomly selected from a pool of potential judges. Table 7.46 contains data for $k = 3$ judges and $N = 5$ subjects. Table 7.47 contains the analysis of variance source table for the data given in Table 7.46. Given the analysis of variance source table in Table 7.47, the intraclass

Table 7.46 Example data for Case 2, Form 1, with $N = 5$ subjects (S) and $k = 3$ judges (A)

	Judge (A)		
Subject (S)	1	2	3
1	12	10	8
2	15	11	7
3	9	9	6
4	6	5	4
5	8	5	5

Table 7.47 Analysis of variance source table for the data given in Table 7.46 with $k = 3$ judges and $N = 5$ subjects

Source	*SS*	*df*	*MS*	*F*
Between S	78.00	4	19.50	
Within S	54.00	10	5.40	
Factor A	40.00	2	20.00	11.43
Error ($A \times S$)	14.00	8	1.75	
Total	132.00	14		

correlation coefficient is

$$r_{\mathrm{I}} = \frac{MS_{\mathrm{BS}} - MS_{\mathrm{A\times S}}}{MS_{\mathrm{BS}} + (a-1)MS_{\mathrm{A\times S}} + \dfrac{a(MS_{\mathrm{A}} - MS_{\mathrm{A\times S}})}{N}}$$
$$= \frac{19.50 - 1.75}{19.50 + (3-1)(1.75) + \dfrac{(3)(20.00 - 1.75)}{5}} = 0.5228 .$$

7.8.2 A Permutation Analysis

Permutation analyses are completely data-dependent and do not depend on random sampling and/or fixed- or random-effects models. For the data given in Table 7.46 for $k = 3$ judges and $N = 5$ subjects there are only

$$M = \left(k!\right)^N = \left(3!\right)^5 = 7{,}776$$

possible, equally-likely arrangements in the reference set of all permutations of the observed data, making an exact permutation analysis possible. If r_{o} denotes the observed value of r_{I}, the exact upper-tail probability value of the observed value of r_{I} is

$$P\left(r_{\mathrm{I}} \geq r_{\mathrm{o}} | H_0\right) = \frac{\text{number of } r_{\mathrm{I}} \text{ values } \geq r_{\mathrm{o}}}{M} = \frac{24}{7{,}776} = 0.0031 .$$

7.8.3 Interclass and Intraclass Linear Correlation

In the special case of $k = 2$ the relationship between the Pearson product-moment (interclass) correlation coefficient and the Pearson intraclass correlation coefficient can easily be demonstrated. Given $k = 2$ judges, the value of the intraclass correlation depends in part upon the corresponding Pearson product-moment correlation, but it also depends upon the differences between the means and standard deviations of the two variables. Thus,

$$r_{\mathrm{I}} = \frac{\left[\left(\sigma_x^2 + \sigma_y^2\right) - \left(\sigma_x - \sigma_y\right)^2\right] r_{xy} - \left(\bar{x} - \bar{y}\right)^2 / 2}{\left(\sigma_x^2 + \sigma_y^2\right) + \left(\bar{x} - \bar{y}\right)^2 / 2} ,$$

Table 7.48 Example bivariate correlation data on $N = 5$ subjects

Subject	Height (x)	Weight (y)
A	1	4
B	2	3
C	3	5
D	4	7
E	5	6

where $\bar{x}$ and $\bar{y}$ denote the means, σ_x^2 and σ_y^2 the variances, and r_{xy} the Pearson product-moment correlation of variables x and y. Thus, for the bivariate data given in Table 7.38 on p. 428, replicated in Table 7.48 for convenience,

$$\bar{x} = 3.00\,, \quad \bar{y} = 5.00\,, \quad \sigma_x = \sigma_y = 1.4142\,, \quad \sigma_x^2 = \sigma_y^2 = 2.00\,,$$

$r_{xy} = +0.80$, and

$$r_{\mathrm{I}} = \frac{\left[2.00 + 2.00 - (1.4142 - 1.4142)^2\right] 0.80 - (3.00 - 5.00)^2/2}{(2.00 + 2.00) + (3.00 - 5.00)^2/2}$$

$$= \frac{1.20}{6.00} = +0.20\,,$$

the same value found with $2N$ pairs of observations.

7.9 Coda

Chapter 7 applied permutation statistical methods to measures of association for two variables at the interval level of measurement. Included in Chap. 7 were discussions of ordinary least squares (OLS) regression, least absolute deviation (LAD regression), multivariate multiple regression, point-biserial correlation, biserial correlation, intraclass correlation, and Fisher's z transform for skewed distributions.

Chapter 8 applies exact and Monte Carlo resampling permutation statistical methods to measures of association for two variables at different levels of measurement, e.g., a nominal-level variable and an ordinal-level variable, a nominal-level variable and an interval-level variable, and an ordinal-level variable and an interval-level variable. Included in Chap. 8 are permutation statistical methods applied to Freeman's θ, Agresti's $\hat{\delta}$, Piccarreta's $\hat{\tau}$, Whitfield's S, Cureton's r_{rb}, Pearson's η^2, Kelley's ϵ^2, Hays' $\hat{\omega}^2$, and Jaspen's multiserial correlation coefficient.

References

1. Barrodale, I., Roberts, F.D.K.: A improved algorithm for discrete ℓ_1 linear approximation. J. Num. Anal. **10**, 839–848 (1973)
2. Barrodale, I., Roberts, F.D.K.: Solution of an overdetermined system of equations in the ℓ_1 norm. Commun. ACM **17**, 319–320 (1974)
3. Bartko, J.J.: The intraclass correlation coefficient as a measure of reliability. Psychol. Rep. **19**, 3–11 (1966)
4. Bartko, J.J.: On various intraclass correlation reliability coefficients. Psychol. Bull. **83**, 762–765 (1976)
5. Berry, K.J., Mielke, P.W.: A Monte Carlo investigation of the Fisher Z transformation for normal and nonnormal distributions. Psychol. Rep. **87**, 1101–1114 (2000)
6. Berry, K.J., Mielke, P.W., Johnston, J.E.: Permutation Statistical Methods: An Integrated Approach. Springer–Verlag, Cham, CH (2016)
7. David, F.N.: Tables of the Distribution of the Correlation Coefficient. Cambridge University Press, Cambridge, UK (1938)
8. Dielman, T.E.: A comparison of forecasts from least absolute and least squares regression. J. Forecasting **5**, 189–195 (1986)
9. Dielman, T.E.: Corrections to a comparison of forecasts from least absolute and least squares regression. J. Forecasting **8**, 419–420 (1989)
10. Dielman, T.E., Pfaffenberger, R.: Least absolute value regression: Necessary sample sizes to use normal theory inference procedures. Dec. Sci. **19**, 734–743 (1988)
11. Endler, J.A., Mielke, P.W.: Comparing entire colour patterns as birds see them. Biol. J. Linn. Soc. **86**, 405–431 (2005)
12. Feinstein, A.R.: Clinical Biostatistics XXIII: The role of randomization in sampling, testing, allocation, and credulous idolatry (Part 2). Clin. Pharmacol. Ther. **14**, 898–915 (1973)
13. Fisher, R.A.: Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika **10**, 507–521 (1915)
14. Fisher, R.A.: Studies in crop variation, I. An examination of the yield of dressed grain from Broadbalk. J. Agric. Sci. **11**, 107–135 (1921)
15. Fisher, R.A.: Statistical Methods for Research Workers, 5th edn. Oliver and Boyd, Edinburgh (1934)
16. Flanagan, J.C.: General considerations in the selection of test items and a short method of estimating the product-moment coefficient from the data at the tails of the distribution. J. Educ. Psych. **30**, 674–680 (1939)
17. Gayen, A.K.: The frequency distribution of the product-moment correlation coefficient in random samples of any size drawn from non-normal universes. Biometrika **38**, 219–247 (1951)
18. Geary, R.C.: Testing for normality. Biometrika **34**, 209–242 (1947)
19. Goodman, L.A.: Measures, models, and graphical displays in the analysis of cross-classified data. J. Am. Stat. Assoc. **86**, 1085–1111 (1991)
20. Haggard, E.A.: Intraclass Correlation and the Analysis of Variance. Dryden, New York (1958)
21. Hotelling, H.: New light on the correlation coefficient and its transforms. J. R. Stat. Soc. Meth **15**, 193–232 (1953)
22. Jeyaratnam, S.: Confidence intervals for the correlation coefficient. Stat. Probab. Lett. **15**, 389–393 (1992)
23. Johnston, J.E., Berry, K.J., Mielke, P.W.: Permutation tests: Precision in estimating probability values. Percept. Motor Skill **105**, 915–920 (2007)
24. Kaufman, E.H., Taylor, G.D., Mielke, P.W., Berry, K.J.: An algorithm and FORTRAN program for multivariate LAD (ℓ_1 of ℓ_2) regression. Computing **68**, 275–287 (2002)
25. Kraemer, H.C.: Improved approximation to the non-null distribution of the correlation coefficient. J. Am. Stat. Assoc. **68**, 1004–1008 (1973)
26. Kraemer, H.C., Thiemann, S.: How Many Subjects? Sage, Newbury Park, CA (1987)
27. Krause, E.F.: Taxicab Geometry. Addison–Wesley, Menlo Park, CA (1975)

28. Liu, W.C., Woodward, J.A., Bonett, D.G.: The generalized likelihood ratio test for the Pearson correlation. Commun. Stat. Simul. C **25**, 507–520 (1996)
29. Mathew, T., Nordström, K.: Least squares and least absolute deviation procedures in approximately linear models. Stat. Probab. Lett. **16**, 153–158 (1993)
30. Matthews, R.: Beautiful, but dangerous. Significance **13**, 30–31 (2016)
31. McGrath, R.E., Meyer, G.J.: When effect sizes disagree: The case of r and d. Psychol. Meth. **11**, 386–401 (2006)
32. McGraw, K.O., Wong, S.P.: Forming inferences about some intraclass correlation coefficients. Psychol. Meth. **1**, 30–46 (1996)
33. Micceri, T.: The unicorn, the normal curve, and other improbable creatures. Psychol. Bull. **105**, 156–166 (1989)
34. Mielke, P.W.: Asymptotic behavior of two-sample tests based on powers of ranks for detecting scale and location alternatives. J. Am. Stat. Assoc. **67**, 850–854 (1972)
35. Mielke, P.W.: Another family of distributions for describing and analyzing precipitation data. J. Appl. Meteor. **12**, 275–280 (1973). [Corrigendum: J. Appl. Meteor. **13**, 516 (1973)]
36. Mielke, P.W., Berry, K.J.: Multivariate multiple regression analyses: A permutation method for linear models. Psychol. Rep. **91**, 3–9 (2002)
37. Mielke, P.W., Berry, K.J.: Multivariate multiple regression prediction models: A Euclidean distance approach. Psychol. Rep. **92**, 763–769 (2003)
38. Mielke, P.W., Berry, K.J.: Permutation Methods: A Distance Function Approach, 2nd edn. Springer–Verlag, New York (2007)
39. Mielke, P.W., Berry, K.J., Landsea, C.W., Gray, W.M.: Artificial skill and validation in meteorological forecasting. Weather Forecast **11**, 153–169 (1996)
40. Mielke, P.W., Berry, K.J., Landsea, C.W., Gray, W.M.: A single-sample estimate of shrinkage in meteorological forecasting. Weather Forecast **12**, 847–858 (1997)
41. Mudholkar, G.S., Chaubey, Y.P.: On the distribution of Fisher's transformation of the correlation coefficient. Commun. Stat. Simul. C **5**, 163–172 (1976)
42. Nunnally, J.C.: Psychometric Theory, 2nd edn. McGraw–Hill, New York (1978)
43. Pearson, E.S.: Some notes on sampling with two variables. Biometrika **21**, 337–360 (1929)
44. Pfaffenberger, R., Dinkel, J.: Absolute deviations curve-fitting: An alternative to least squares. In: David, H.A. (ed.) Contributions to Survey Sampling and Applied Statistics, pp. 279–294. Academic Press, New York (1978)
45. Pillai, K.C.S.: Confidence interval for the correlation coefficient. Sankhyā **7**, 415–422 (1946)
46. Robinson, W.S.: The statistical measurement of agreement. Am. Sociol. Rev. **22**, 17–25 (1957)
47. Rousseeuw, P.J.: Least median of squares regression. J. Am. Stat. Assoc. **79**, 871–880 (1984)
48. Ruben, H.: Some new results on the distribution of the sample correlation coefficient. J. R. Stat. Soc. **28**, 513–525 (1966)
49. Samiuddin, M.: On a test for an assigned value of correlation in a bivariate normal distribution. Biometrika **57**, 461–464 (1970)
50. Shrout, P.E., Fleiss, J.L.: Intraclass correlations: Uses in assessing rater relaibility. Psychol. Bull. **86**, 420–428 (1979)
51. Sitgreaves, R.: Review of "Intraclass Correlation and the Analysis of Variance" by E. A. Haggard. J. Am. Stat. Assoc. **55**, 384–385 (1960)
52. Taylor, L.D.: Estimation by minimizing the sum of absolute errors. In: Zarembka, P. (ed.) Frontiers in Econometrics, pp. 169–190. Academic Press, New York (1974)
53. Thompson, D.: Volcano Cowboys. St. Martin's Press, New York (2000)
54. von Eye, A., Mun, E.Y.: Analyzing Rater Agreement. Lawrence Erlbaum, Mahwah, NJ (2005)
55. Wilson, H.G.: Least squares versus minimum absolute deviations estimation in linear models. Dec. Sci. **9**, 322–325 (1978)
56. Winer, B.J.: Statistical Principles in Experimental Design, 2nd edn. McGraw–Hill, New York (1971)
57. Wong, S.P., McGraw, K.O.: Confidence intervals and F tests for intraclass correlations based on three-way random effects models. Educ. Psychol. Meas. **59**, 270–288 (1999)