include several that fall in the category of a conflict measure rather than an exposure measure. It is recommended that future research closely examine the exposure versus conflict issue as well as the sensitivity of the resulting accident rate expressions to typical countermeasures.

REFERENCES

1. F. M. Council, J. R. Steward, D. W. Reinfurt, and W. W. Hunter. *Exposure Measures for Evaluating Highway Safety Issues*, Vol. 1: *Final Report.* Report FHWA/RD-83/088. University of North Carolina Highway Safety Research Center, Chapel Hill, N.C., 1983.
2. M. Plass. *Evaluation of Exposure-Based Accident Rates.* Department of Civil and Environmental Engineering, University of Wisconsin-Madison, Jan. 1985.
3. G. Keppel. *Design and Analysis: A Researcher's Handbook*, 2nd ed. Prentice-Hall, Englewood Cliffs, N.J., 1982.
4. E. Hauer. Traffic Conflicts and Exposure. *Accident Analysis and Prevention*, Vol. 14, No. 5, 1982, pp. 359–364.
5. W. D. Glauz and D. J. Migletz. *NCHRP Report 219: Application of Traffic Control Analysis at Intersections.* TRB, National Research Council, Washington, D.C., 1980.
6. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985.

Publication of this paper sponsored by Committee on Methodology for Evaluating Highway Improvements.

Demonstration of Regression Analysis with Error in the Independent Variable

RICHARD M. WEED AND RICARDO T. BARROS

Regression analysis is frequently used in the engineering field to develop mathematical models for a wide variety of applications. Of the several assumptions upon which regression theory is based, one of the most fundamental is that the *X*-values are known exactly and that any error is associated only with the *Y*-measurements. Because this is not the case for many engineering applications, a study was conducted (*a*) to determine the magnitude of this problem and (*b*) to develop and test a software package that incorporates a theoretical solution found in the literature. Computer simulation is used to demonstrate both the seriousness of the problem and the effectiveness of the solution. An example based on early-strength tests of concrete is presented.

Many engineering applications require the development of a mathematical model (equation) to characterize some physical relationship. Examples include those shown in Table 1.

In the first example, the objective is a reliable early predictor of the 28-day strength of concrete, a measure upon which many acceptance procedures are based. The objective of the second example is to replace a costly and time-consuming subjective rating procedure with a simple mechanical device. In the third example, a relationship is sought that will become an integral part of a pavement management system.

New Jersey Department of Transportation, 1035 Parkway Avenue, Trenton, N.J. 08625.

The variable to be predicted or estimated is placed on the *Y*-axis and an equation of the form $y = f(x)$ is desired. The equation may be linear, quadratic, exponential, or any other appropriate form. The analyst, from his understanding of the physical process, will often know the correct form in advance. In other cases, it may be necessary to let the data dictate the form.

The desired relationship is often derived empirically from a set of *X*,*Y*-data values by using the technique of least squares (*1*) as shown in Figure 1. The procedure, invisible to the analyst when executed by a computer program, consists of solving for

TABLE 1 PHYSICAL RELATIONSHIPS CHARACTERIZED BY MATHEMATICAL MODELS

Characteristic of Interest	*X*-Data (Independent Variable)	*Y*-Data (Dependent Variable)
Compressive strength of concrete	Seven-day test results	Twenty-eight-day test results
Rating of highway pavement serviceability	Output of mechanical roughness-measuring device	Average rating of a team of panelists
Rating of highway pavement serviceability	Cumulative axle loads	Current rating of serviceability

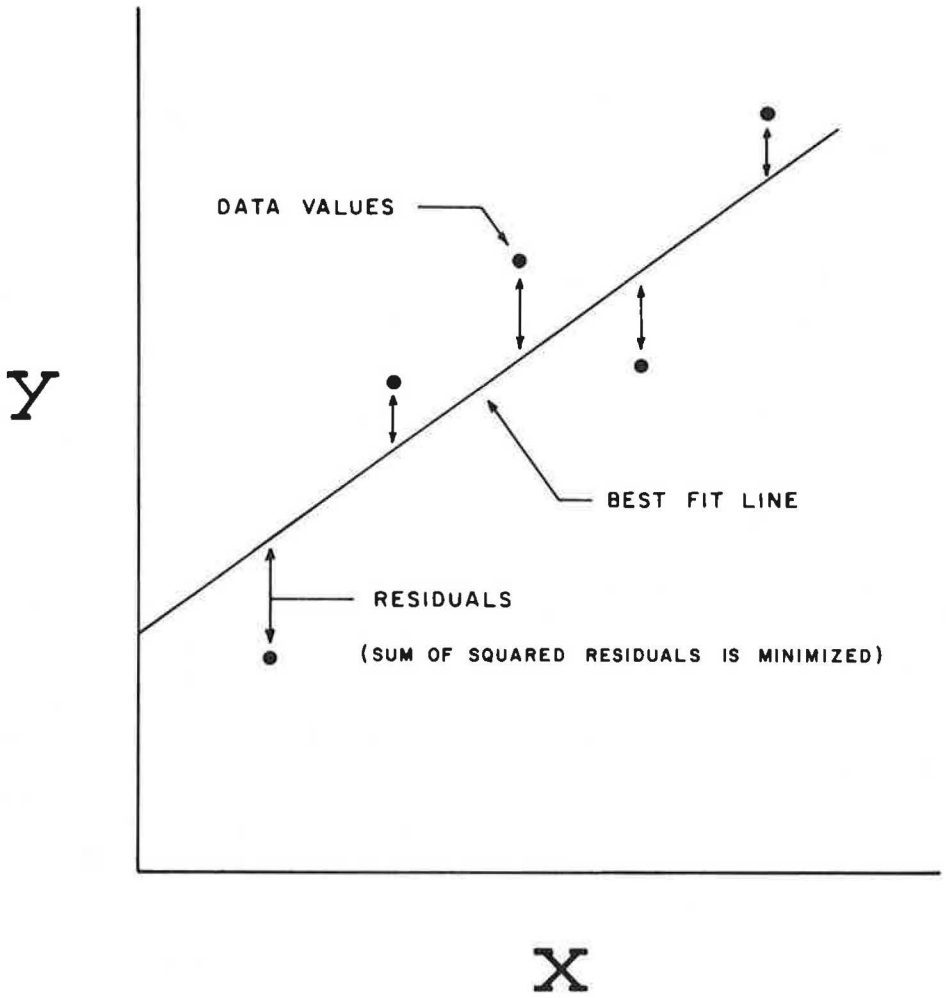

FIGURE 1 Concept of the ordinary least-squares technique.

the line that best fits the data. When ordinary least squares is used, the best fit is defined as that line of the chosen form that minimizes the sum of the squared residuals in a direction parallel to the *Y*-axis.

In carrying out this procedure, the user makes several theoretical assumptions, one of the most fundamental of which is that the *X*-values are known exactly and that any error of measurement is associated only with the *Y*-values. Because it is often impossible or impractical to achieve this idealized condition in practice, a study (*2*) was undertaken (*a*) to investigate the effect of failing to satisfy this assumption and (*b*) to develop and test a software package that incorporates a theoretical procedure for dealing with *X*-error (*3*). Linear models are addressed because only the linear solution of the *X*-error problem has been published to date.

USE OF COMPUTER SIMULATION

In order to demonstrate the extent of the *X*-error problem and the effectiveness of the solution, a method was required to observe and quantify the accuracy and precision of the regression estimates. This can readily be accomplished with computer simulation by performing the following steps:

1. Randomly generate a bivariate normal *X*,*Y*-data set with known regression (population) parameters:
   (*a*) intercept (β_0),
   (*b*) slope (β_1), and
   (*c*) residual error (σ_{yx}).
2. Include a fixed amount of *X*-error, either in absolute terms or as a percentage of σ_{yx}.
3. Use the randomly generated data to estimate the regression parameters:
   (*a*) intercept (B_0),
   (*b*) slope (B_1), and
   (*c*) residual error (S_{yx}).
4. Repeat the entire process many times in order to compare the distributions of the regression estimates with the known parameters. Ideally, the sampling distributions of the estimates should be centered on the true population parameters and have relatively narrow dispersions.

This technique can be used to provide a very dramatic demonstration of the bias introduced by error in the *X*-variable and the conditions that accentuate it. It will also be used to demonstrate the effectiveness of the procedure developed to overcome this problem.

TABLE 2 EXAMPLES OF BIAS INTRODUCED BY THE PRESENCE OF X-ERROR

Parameter	True Value	Regression Estimates Obtained by Ordinary Least Squares for Selected Levels of X-Error[a]				
		0	25	50	75	100
Intercept	100	100.14	108.11	129.25	160.81	200.23
Slope	10	10.00	9.84	9.41	8.78	8.00
Residual error	5	4.93	13.21	24.62	35.08	44.79

NOTE: Results obtained by computer simulation with 1,000 replications of 30 data points spanning the range between approximately $X = 30$ and $X = 70$.

[a]X-error is measured as the ratio σ_{xx}/σ_{yx}, expressed as a percentage, in which σ_{xx} represents the error in individual X-measurements.

TABLE 3 EFFECT OF SLOPE ON THE DEGREE OF BIAS INTRODUCED

Parameter	True Value	Regression Estimates Obtained by Ordinary Least Squares for Fixed X-Error[a] and Selected Levels of Slope				
		0.0	0.5	1.0	5.0	10.0
Intercept	100	99.94	105.14	109.95	150.26	200.23
Slope	–[b]	0.00	0.40	0.80	3.99	8.00
Residual error	5	4.95	5.44	6.69	22.58	44.79

NOTE: Results obtained by computer simulation with 1,000 replications of 30 data points spanning the range between approximately $X = 30$ and $X = 70$.

[a]The level of X-error is fixed at 100 percent ($\sigma_{xx} = \sigma_{yx}$ in which σ_{xx} represents the error in individual X-measurements).
[b]Variable (values given in column headings).

DEMONSTRATION OF THE PROBLEM

Table 2 has been prepared with dimensionless data to demonstrate the detrimental effect that even a moderate amount of X-error can have under certain conditions. It may be observed that when there is no X-error, the estimated values (averages for 1,000 replications) of all three regression parameters are extremely close to the true population values. When the amount of X-error is as little as 25 percent of the Y-error (σ_{yx}), it may be seen that a substantial amount of bias has been introduced in the estimates of both the intercept and the residual error. As the degree of X-error increases, all three regression parameters begin to show considerable bias. When the X-error and the Y-error are approximately equal—a fairly common situation in actual practice—the regression estimates differ substantially from the true population parameters.

It should be noted that this example was chosen to dramatize the potentially serious nature of the X-error problem. Although the three population parameters ($\beta_0 = 100$, $\beta_1 = 10$, $\sigma_{yx} = 5$) are not extreme in any sense, the effect is pronounced because the slope is fairly steep. Figure 2 presents a conceptual illustration of the effect of the slope in translating X-error into apparent Y-error, error that the ordinary least-squares procedure attributes solely to the Y-variable. The examples in Table 3

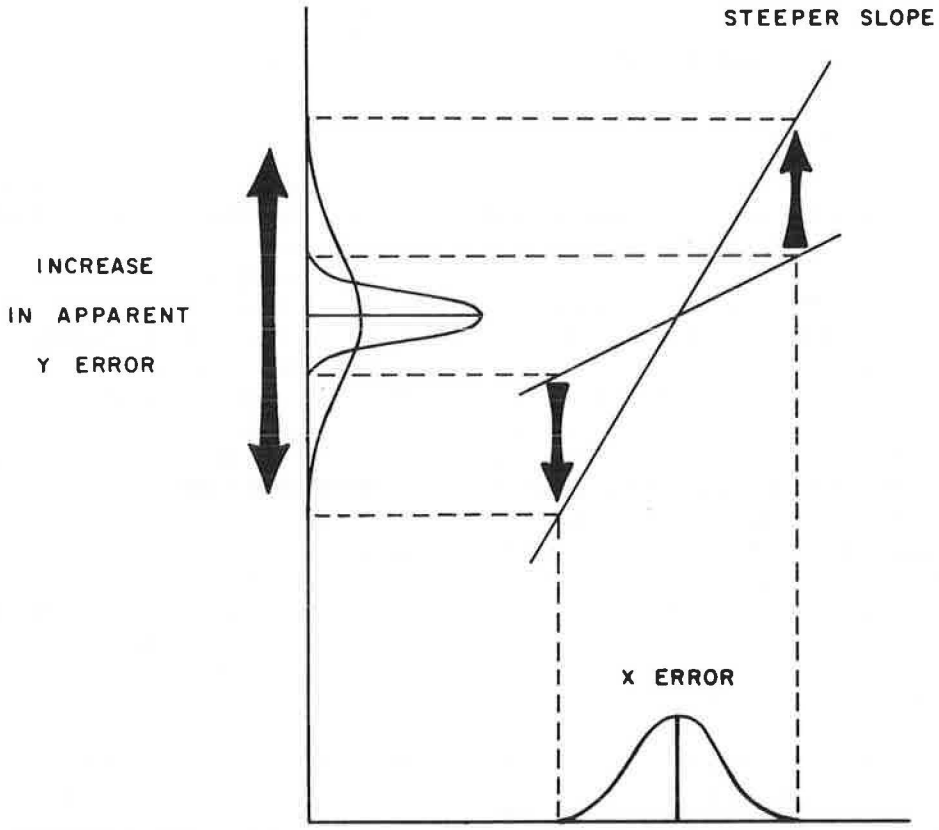

FIGURE 2 Effect of slope on the translation of X-error into apparent Y-error.

further demonstrate this effect. Viewed collectively, the examples in Tables 2 and 3 provide an empirical indication of the conditions that tend to influence the magnitude of the *X*-error problem: the ratio of *X*-error to *Y*-error and the slope of the regression line.

MANDEL'S SOLUTION

A theoretically derived procedure for avoiding the bias in the regression estimates due to *X*-error has been published by Mandel (*3*). Use of the procedure requires one additional bit of information that is usually readily available or readily obtainable: the ratio of the variances associated with the *X*- and *Y*-measurements. Although the mathematical procedure is somewhat involved, the concept is easy to visualize. Ordinary least squares minimizes the sum of the squared residuals in a direction parallel to the *Y*-axis. In the presence of *X*-error, the minimization process is performed by Mandel's procedure in a direction oblique to the *X*- and *Y*-axes, the exact angle being determined primarily by the relative magnitude of the *X*- and *Y*-error.

In order to test the effectiveness of Mandel's method, it was applied to the same data sets used to develop Table 2. The results are reported in Table 4 and the values from Table 2 are repeated for ease of comparison. It can be seen from Table 4 that Mandel's method is extremely effective in removing the bias that exists when an application with *X*-error is analyzed by ordinary least squares. Its only discernible weakness in this example is a possible small downward bias of the estimate of the intercept when the *X*-error is quite large.

To judge whether this apparent bias was real, the simulation program was modified to print out a histogram and elementary statistics for 1,000 intercept estimates. The *X*-error was held constant at 100 percent of σ_{yx}. Although not strictly applicable because the distribution of intercept estimates was somewhat skewed, a *t*-test indicated that the average intercept of 93.06 was highly significantly different ($\alpha < 0.001$) from the true value of 100.0. Although it is not obvious from the results in Table 4, a similar test suggests that the slope estimates may also be biased to a very small degree. Consequently, although Mandel's method appears to be very effective and is far superior to ordinary least squares, it must be concluded that it is not totally unbiased in all cases.

Figure 3 shows in a graphical way the effects that have been observed in Table 4. The distributions shown were drawn from histograms generated by the same simulation programs used to develop Tables 2–4. Mandel's method can be seen to be essentially unbiased in that the means of the distributions generated by that method are very close to the true population parameters. In marked contrast, the distributions produced by ordinary least squares are shifted substantially away from the true parameters. Another important observation in Figure 3 is that the distributions for the intercept and the slope obtained with Mandel's method are only slightly more dispersed than those obtained by ordinary least squares. This indicates that a substantial gain in accuracy has been achieved with only a slight loss of precision. For the residual error, Mandel's method is both more accurate and more precise.

An interesting feature of Mandel's method is that, unlike the least-squares technique, the same regression line will be obtained regardless of which variable is considered to be independent (*X*) and which is dependent (*Y*). Furthermore, the degree of uncertainty associated with predictions made by this method is the same for either choice of variables (*3*, p.9). This property conveniently avoids a controversial aspect of the calibration application, the need to work backward through the regression procedure to estimate what value of *X* gave rise to an observed value of *Y*.

Another series of computer simulation tests was performed by using the appropriate procedures for computing interval estimates for the intercept, slope, and σ_{yx}. A level of confidence of $1 - \alpha = 0.95$ was selected and the number of times that the interval estimate actually contained the true population parameter was counted. For 1,000 replications, the empirically observed results should fall within the range of approximately $0.95 \pm 2\{[(0.95)(0.05)]/1,000\}^{1/2} = 0.95 \pm 0.014$ when the interval estimation process is working properly. It can be seen from the results in Table 5 that, even for small amounts of *X*-error, the interval estimates computed by ordinary least squares contain the population parameters substantially less often than desired. In contrast, all of the interval estimates computed by Mandel's method are satisfactory.

TABLE 4 COMPARISON OF MANDEL'S METHOD WITH ORDINARY LEAST SQUARES

Parameter	True Value	Method	Regression Estimates Obtained for Selected Levels of *X*-Error[a] 0	25	50	75	100
Intercept	100	Mandel	100.18	100.19	98.34	96.69	93.06
		OLS	100.14	108 11	129.25	160.81	200.23
Slope	10	Mandel	10.00	9.99	10.03	10.06	10.14
		OLS	10.00	9.84	9.41	8.78	8.00
Residual error	5	Mandel	4.93	4.94	4.97	4.94	4.97
		OLS	4.93	13.21	24.62	35.08	44.79

NOTE: Results obtained by computer simulation with 1,000 replications of 30 data points spanning the range between approximately $X = 30$ and $X = 70$. OLS = ordinary least squares.

[a] *X*-error is measured as the ratio σ_{xx}/σ_{yx}, expressed as a percentage, in which σ_{xx} represents the error in individual *X*-measurements.

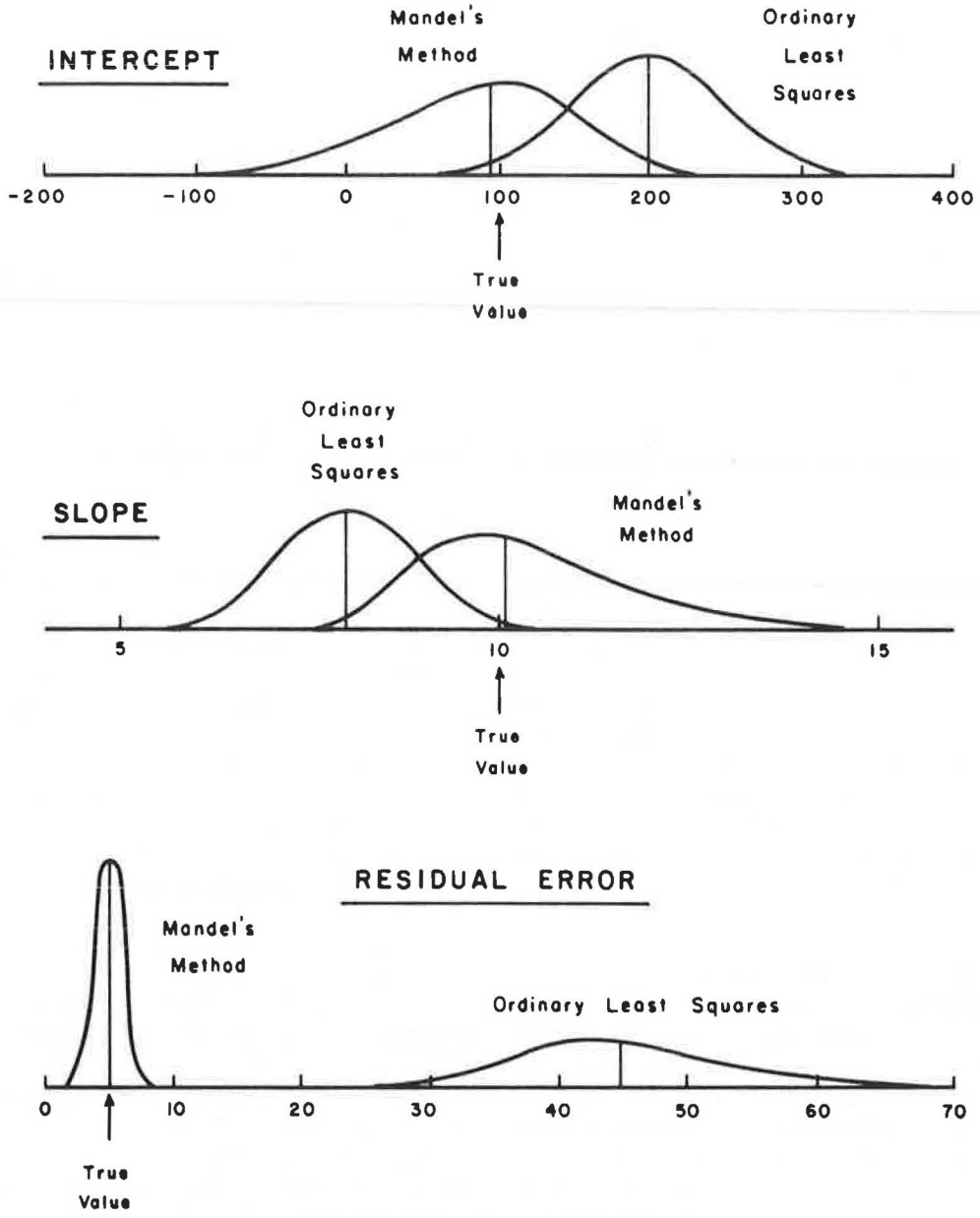

FIGURE 3 Comparison of distributions of regression estimates.

TABLE 5 EFFECT OF *X*-ERROR ON INTERVAL ESTIMATES

Parameter	Desired Confidence Level	Method	Empirically Observed Confidence Levels at Selected Levels of *X*-Error[a]				
			0	25	50	75	100
Intercept ($\beta_0 = 100$)	0.95	Mandel	0.950	0.938	0.954	0.939	0.953
		OLS	0.951	0.898	0.768	0.548	0.321
Slope ($\beta_1 = 10$)	0.95	Mandel	0.948	0.941	0.956	0.948	0.952
		OLS	0.946	0.889	0.755	0.534	0.301
Residual error ($\sigma_{yx} = 5$)	0.95	Mandel	0.955	0.956	0.963	0.947	0.959
		OLS	0.955	0.0	0.0	0.0	0.0

NOTE: Results obtained by computer simulation with 1,000 replications of 30 data points spanning the range between approximately $X = 30$ and $X = 70$. OLS = ordinary least squares.

[a] *X*-error is measured as the ratio σ_{xx}/σ_{yx}, expressed as a percentage, in which σ_{xx} represents the error in individual *X*-measurements.

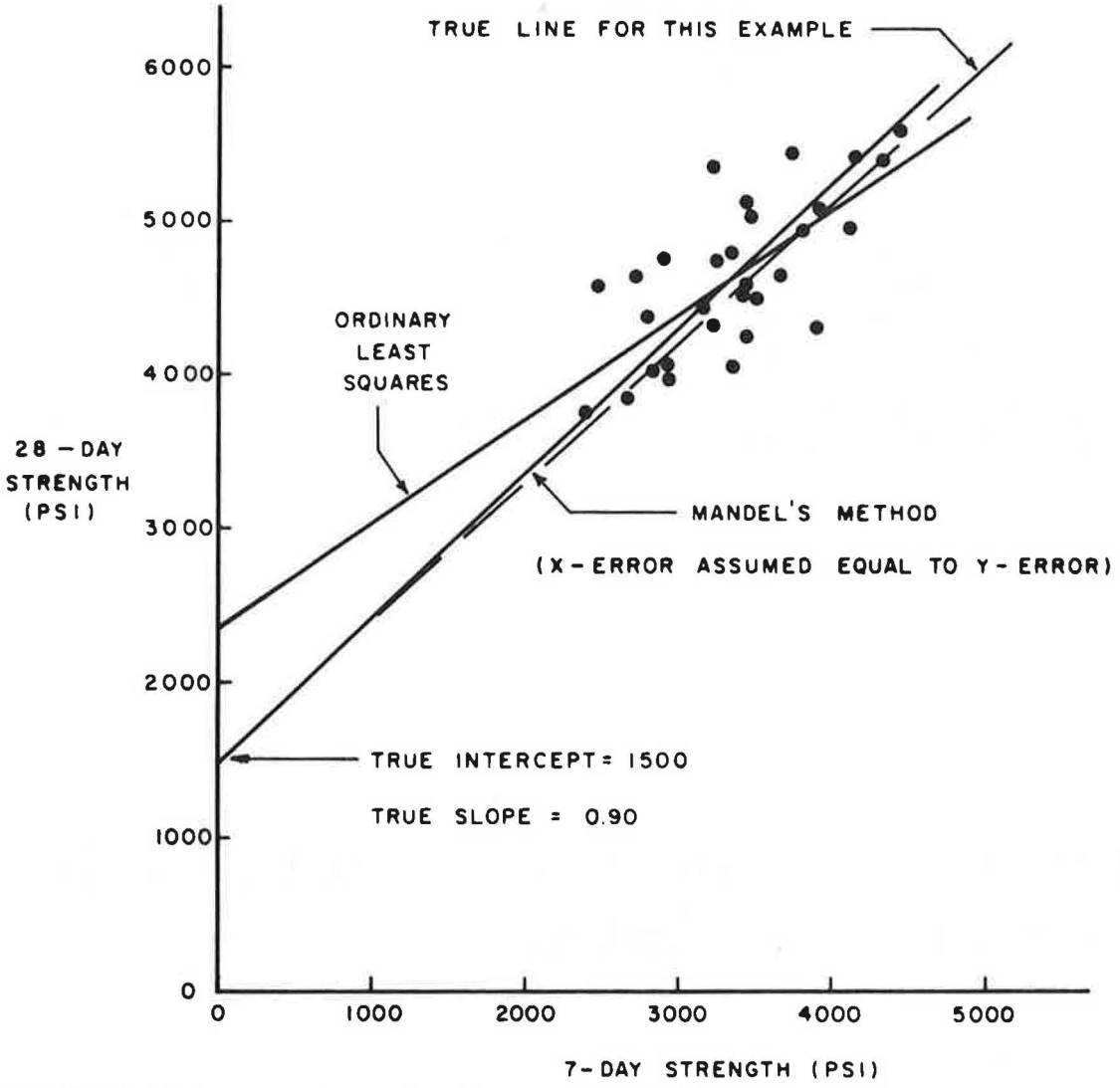

FIGURE 4 Typical regression results with concrete strength data.

EXAMPLE BASED ON CONCRETE STRENGTH DATA

The following example is based on concrete strength data collected from a construction project in New Jersey. It is a contrived example in that the data set that is used was randomly generated from a population having the same statistical parameters as those observed in the field. This approach provides a known control against which the results obtained by the two methods may be compared. Otherwise, it could only be observed that the results obtained by the two methods were distinctly different and it would not be known how close either one came to estimating the true population parameters.

In order to use Mandel's procedure, the ratio of the X- and Y-error variances must be known or assumed. Because both the 7-day and 28-day strengths are at relatively high levels, it has been assumed that the measurement error is the same for both sets of data. Therefore, the variance ratio to be entered into the Mandel procedure is 1.0.

The data points and regression results are shown in Figure 4. These results are typical in that they are examples of central values of the distributions shown in Figure 3. Like the dimensionless examples in Table 4, ordinary least squares has produced a considerably biased estimate, whereas Mandel's method has produced an estimate very close to the true location of the line.

SUMMARY AND CONCLUSIONS

Regression analysis is frequently used in the engineering profession to develop mathematical models for many different applications. In regression theory the X-values are assumed to be known without error, a requirement that often cannot be met. Computer simulation was used to demonstrate that under certain relatively common conditions, regression estimates obtained by ordinary least squares can be seriously in error.

Conditions that appear to warrant concern are (*a*) *X*-error approaching the level of *Y*-error combined with moderate degrees of slope or (*b*) lesser degrees of *X*-error combined with greater degrees of slope.

It was also demonstrated by computer simulation that Mandel's method, which might be termed "oblique least squares" because of the manner in which it minimizes the sum of squared residuals, is extremely effective at removing most of the bias introduced by error in the *X*-variable. Figure 3 and Tables 4 and 5 clearly show that, in general, Mandel's method provides substantially more accurate results than ordinary least squares and Figure 4 illustrates this fact with a specific example based on concrete strength tests. The complete theoretical development, along with a more quantitative guideline to determine when it is advisable to use it, is contained in the original source document (*3*). The FORTRAN coding necessary to apply the procedure is contained in the project report (*2*).

REFERENCES

1. J. Neter, W. Wasserman, and M. Kutner. *Applied Linear Regression Models*. Richard D. Irwin, Homewood, Ill., 1983.
2. R. T. Barros and R. M. Weed. *Applied Regression in the Presence of X Error*. FHWA, U.S. Department of Transportation, 1987.
3. J. Mandel. Fitting Straight Lines When Both Variables Are Subject to Error. *Journal of Quality Technology*, Vol. 16, No. 1, 1984, pp. 1–14.

Publication of this paper sponsored by Committee on Methodology for Evaluating Highway Improvements.

Validation of a Nonautomated Speed Data Collection Methodology

FRED R. HANSCOM

The objective of this research was to develop a validated spot speed study procedure that does not rely on automated equipment. The field study procedure applied a variety of speed collection techniques and compared results against baseline speeds obtained with reliable pavement instrumentation. A recommended manual-timing technique was based on observed accuracies with various vehicle-selection strategies, site conditions, sample sizes, observation period lengths, and observer characteristics.

The conduct of spot speed studies with nonautomated equipment involves a variety of methodological considerations (*1*). Although such studies have long been used in traffic engineering, a number of factors have hampered their valid application (*2*). Among these factors are observer vehicle-selection bias (e.g., the human ability to select a truly random sample), impact of vantage point (e.g., cosine error associated with radar measurement), technique reliability (e.g., stopwatch timing measurement error), and observer human factors (e.g., experience, fatigue).

The objective of this research was to address the effects of the foregoing factors in order to develop a spot speed data collection procedure that does not rely on automated equipment. The field study procedure involved applying a variety of speed collection techniques and comparing results against baseline speeds obtained with reliable pavement instrumentation. A recommended manual timing technique was based on achieved accuracies with various vehicle-selection strategies, site conditions, sample sizes, observation period lengths, and observer characteristics.

Transportation Research Corporation, 2710 Ridge Road, Haymarket, Va. 22069.

VEHICLE-SELECTION STRATEGIES

Basic Application

Specific techniques were evaluated that controlled observer bias in selecting vehicles for speed measurement. Thus, two speed collection methods (radar and manual timing) were applied by using the following vehicle-selection strategies: