PROPERTIES OF THE CORRECTION FOR ATTENUATION ESTIMATOR

Wayne A. Fuller, Iowa State University
Michael A. Hidiroglou, Statistics Canada

ABSTRACT

The limiting distribution of the regression coefficients calculated from a correlation matrix that has been corrected for attenuation is obtained. Methods of estimating the covariance matrix of the vector of regression coefficients are presented. Nonnormal regression variables and nondiagonal error matrices are considered. The procedures are illustrated with data on the socioeconomic career.

1. INTRODUCTION

The effect of measurement error upon estimated regression coefficients has long been recognized. Cochran [5], Johnston [9], Walker and Lev [16] and Wiley [18] are recent references reporting the distortions that are introduced into standard regression statistics when the independent variables are measured with error. In a regression with a single independent variable the regression coefficients, on average, are reduced in absolute value, attenuated, when compared to those computed in the absence of measurement error. The same is true of the correlation coefficients.

In some areas it is possible to obtain good estimates of the ratio of the measurement error variance to the total variance. If the measurement errors in different independent variables are uncorrelated, the estimated variance ratios can be used to adjust the observed correlation matrix to construct an estimate of the correlation matrix one would obtain in the absence of measurement errors. The resulting estimated correlation matrix is said to have been corrected for attenuation. Regression equations can then be estimated from the correlation (or covariance) matrix corrected for attenuation. Although the method has been extensively used in the social sciences, little discussion of the sampling properties of the estimators is available (see Bohrnstedt and Carter [3]).

In this paper we derive the limiting distribution for the correction for attenuation estimator for both the uncorrelated and correlated measurement error cases. We also demonstrate how the standard error of the regression coefficients can be estimated when the error and (or) the true values have an arbitrary distribution with finite fourth moments.

The distributional results are illustrated using the causal chain model for the socioeconomic career discussed by Featherman [6] and Kelley [12].

2. MODEL AND ESTIMATION

We write the model as

$$\underset{\sim}{Y} = \underset{\sim}{x}\,\underset{\sim}{\beta} + \underset{\sim}{e}$$
$$\underset{\sim}{X} = \underset{\sim}{x} + \underset{\sim}{u}\,, \qquad (2.1)$$

where $\underset{\sim}{Y}$ is an $n \times 1$ vector, $\underset{\sim}{x}$ is an $n \times k$ matrix, and $\underset{\sim}{\beta}$ is a $k \times 1$ vector. The vector $\underset{\sim}{Y}$ and the matrix $\underset{\sim}{X}$ are observed and an estimator of $\underset{\sim}{\beta}$ is desired. The matrix $\underset{\sim}{u}$ is the matrix of measurement errors. We shall utilize the following assumptions:

(i) The vectors of errors $(e_t, \underset{\sim}{u}_t)$, $t=1, 2, \ldots$, where $\underset{\sim}{u}_t$ is the t^{th} row of $\underset{\sim}{u}$ are distributed as normal independent random variables with zero mean and covariance matrix

$$\begin{pmatrix} \sigma_e^2 & \underset{\sim}{\Sigma}_{eu} \\ \underset{\sim}{\Sigma}_{ue} & \underset{\sim}{\Sigma}_{uu} \end{pmatrix} = \operatorname{diag}(\sigma_e^2, \sigma_{u_1}^2, \sigma_{u_2}^2, \ldots, \sigma_{u_k}^2) .$$

(ii) The distribution of $(e_j, \underset{\sim}{u}_j)$ is independent of that of $\underset{\sim}{x}_t$ for all t, j where $\underset{\sim}{x}_t$ is the t^{th} row of $\underset{\sim}{x}$.

(iii) The $\underset{\sim}{x}_t$, $t = 1,2,\ldots,n$, are distributed as normal independent random variables with mean $\underset{\sim}{0}$ and nonsingular covariance matrix $\underset{\sim}{\Sigma}_{xx}$.

The reader will note that we have lost no generality in assuming the mean of the $\underset{\sim}{x}_t$ to be zero. If the mean is unknown we make an orthogonal transformation to reduce the problem to the stated form. In practice one uses the corrected sums of squares and products in the analysis when the mean in unknown. If an independent variable is measured without error, then $\sigma_{u_i}^2 = 0$ for that variable.

Since $\underset{\sim}{x}_t$ and $\underset{\sim}{u}_t$ are normally distributed it follows that $\underset{\sim}{X}_t = \underset{\sim}{x}_t + \underset{\sim}{u}_t$, $t = 1,2,\ldots,n$ are distributed as normal independent random variables with mean zero and nonsingular covariance matrix $\underset{\sim}{\Sigma}_{XX} = \underset{\sim}{\Sigma}_{xx} + \underset{\sim}{\Sigma}_{uu}$. It is also assumed that:

(iv) The ratios λ_i, $i = 1,2,\ldots,k$, of error variance $\sigma_{u_i}^2$ to total variance $\sigma_{X_i}^2$, where $\sigma_{u_i}^2$ is the i^{th} diagonal element of $\underset{\sim}{\Sigma}_{uu}$ and $\sigma_{X_i}^2$ is the diagonal element of $\underset{\sim}{\Sigma}_{XX}$, are known.

The quantity $(1-\lambda_i)$ is called the reliability of the i^{th} variable. We denote the diagonal matrix of ratios by

$$\underset{\sim}{\Lambda} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k) . \qquad (2.2)$$

We define the correction for attenuation estimator of $\underset{\sim}{\beta}$ by

$$\hat{\underset{\sim}{\beta}} = \hat{\underset{\sim}{H}}^{-1}(n^{-1}\,\underset{\sim}{X}'\,\underset{\sim}{Y}) , \qquad (2.3)$$

where

$$\hat{\mathbf{H}} = \begin{cases} \frac{1}{n}\mathbf{X}'\mathbf{X} - \mathbf{D}\mathbf{\Lambda}\mathbf{D} & , \text{ if } \hat{f} \geq 1 + n^{-1} \\ \frac{1}{n}\mathbf{X}'\mathbf{X} - (f - n^{-1})\mathbf{D}\mathbf{\Lambda}\mathbf{D} , & \text{ if } \hat{f} < 1 + n^{-1} , \end{cases}$$

$$\mathbf{D} = \mathrm{diag}(s_{X_1}, s_{X_2}, \ldots, s_{X_k}) ,$$

$$s^2_{X_i} = n^{-1} \sum_{t=1}^{n} X^2_{ti} ,$$

$\hat{f}$ is the smallest root of

$$|\mathbf{M} - f\mathbf{T}\mathbf{G}\mathbf{T}| = 0 ,$$

$$\mathbf{M} = \frac{1}{n} \begin{pmatrix} \mathbf{Y}'\mathbf{Y} & \mathbf{Y}'\mathbf{X} \\ \mathbf{X}'\mathbf{Y} & \mathbf{X}'\mathbf{X} \end{pmatrix} ,$$

$$\mathbf{T} = \begin{bmatrix} s_Y & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{bmatrix} ,$$

$$s^2_Y = n^{-1} \sum_{t=1}^{n} Y^2_t \quad ,$$

$$\mathbf{G} = \begin{cases} \mathrm{diag}(0, \lambda_1, \lambda_2, \ldots, \lambda_k) & \text{if the reliability of } Y \text{ is unknown} \\ \mathrm{diag}(\lambda_{ee}, \lambda_1, \lambda_2, \ldots, \lambda_k) & \text{if the reliability of } Y \text{ is known and denoted by } \lambda_{ee} . \end{cases}$$

The slight modification introduced by the calculation of $\hat{f}$ guarantees that the matrix $\hat{\mathbf{H}}$ to be inverted is always positive definite, and that the estimated covariance matrix of the true variables is positive definite. In practice, if one obtains a small $\hat{f}$ one should investigate the hypothesis that the covariance matrix $\mathbf{\Sigma}_{xx}$ is singular by computing the smallest root of

$$|n^{-1}\,\mathbf{X}'\mathbf{X} - \ell\,\mathbf{D}\mathbf{\Lambda}\mathbf{D}| = 0 .$$

If $\hat{\ell}$ is not significantly different from one, it may be desirable to modify the model by reducing the dimension of $\mathbf{X}$. By the results of Fuller [7], the distribution of $(n\text{-}k)\hat{f}$ is approximately that of a chi-square random variable with n-k degrees of freedom when the rank of $(\mathbf{x}{:}\mathbf{y})'(\mathbf{x}{:}\mathbf{y})$, where $\mathbf{y} = \mathbf{x}\boldsymbol{\beta}$, is k and the reliability of Y is known. Similarly, $(n\text{-}k\text{+}1)\hat{\ell}$ is approximately distributed as a chi-square random variable with n-k+1 degrees of freedom when the rank of $\mathbf{x}'\mathbf{x}$ is k-1 .

<u>Theorem 1</u>: Let model (2.1) and assumptions (i) through (iv) hold. Then

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{\Sigma}^{-1}_{xx}\,\mathbf{C}\,\mathbf{\Sigma}^{-1}_{xx}) ,$$

where the ijth element of the matrix $\mathbf{C}$ is

$$c_{ij} = \sigma_{X_iX_j}\left(\sigma^2_v - 2\lambda^2_i\beta^2_i\sigma^2_{X_i} - 2\lambda^2_j\beta^2_j\sigma^2_{X_j} + 2\lambda_i\lambda_j\beta_i\beta_j\sigma_{X_iX_j}\right) + \lambda_i\lambda_j\beta_i\beta_j\sigma^2_{X_i}\sigma^2_{X_j}$$

and $\sigma^2_v = \sigma^2_e + \sum_{i=1}^{k} \beta^2_i\sigma^2_{u_i}$.

Proofs of the theorems may be obtained by writing the authors for the complete manuscript.

The covariance matrix of $\hat{\boldsymbol{\beta}}$ is estimated by replacing the parameters by their estimates, where σ^2_v is estimated by

$$s^2_v = \frac{1}{n\text{-}k} \sum_{t=1}^{n} (Y_t - \mathbf{X}_t\boldsymbol{\beta})^2 ,$$

$\hat{\mathbf{H}}$ is an estimator of $\mathbf{\Sigma}_{xx}$, $n^{-1}\,\mathbf{X}'\mathbf{X}$ furnishes estimators of $\sigma_{X_iX_j}$, and $\mathbf{X}_t$ is the tth row of the matrix $\mathbf{X}$.

In the computations sums of squares corrected for the mean will typically be used throughout. If an intercept term is computed for the regression,

$$\tilde{\beta}_0 = \bar{Y} - \bar{\mathbf{X}}'\tilde{\boldsymbol{\beta}} ,$$

the variance of the estimated intercept can be estimated by

$$\hat{V}\{\tilde{\beta}_0\} = n^{-1}\, s^2_v + \bar{\mathbf{X}}'\hat{\mathbf{H}}^{-1}\,\hat{\mathbf{C}}\hat{\mathbf{H}}^{-1}\,\bar{\mathbf{X}} ,$$

where $\bar{\mathbf{X}}' = (\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k)$ and $\hat{\mathbf{C}}$ is the estimator of $\mathbf{C}$.

The form of the covariance matrix obtained in Theorem 1 was a function of the moment properties of the normal distribution. However, the fact that the estimator converged in distribution to a normal random variable required only independence of the observations and the existence of certain moments. Therefore, we can extend the procedure to nonnormal distributions. We also relax the assumption that the covariance matrix of the measurement errors is diagonal. We make the assumptions:

(v) The vectors $(e_t, \mathbf{u}_t, \mathbf{x}_t)$, t = 1,2,..., are independently and identically distributed with

$$E\{e_t, \mathbf{u}_t\} = \mathbf{0}$$

$$E\{\mathbf{x}_t\} = \boldsymbol{\mu}$$

$$E\{(e_t, \mathbf{u}_t)'(e_t, \mathbf{u}_t)\} = \mathbf{\Sigma},$$

$$E\{(\mathbf{x}_t - \boldsymbol{\mu})'(\mathbf{x}_t - \boldsymbol{\mu})\} = \mathbf{\Sigma}_{xx}$$

$$E\{\mathbf{x}'_t\,(e_t, \mathbf{u}_t)\} = \mathbf{0},$$

and finite fourth moments, where $\mathbf{\Sigma}_{xx}$ is nonsingular.

(vi) The matrices Λ_{eu} and Λ_{uu} are known, where

$$G^{\dagger} = \mathcal{D}^{-1} \Sigma \mathcal{D}^{-1} = \begin{pmatrix} \lambda_{ee} & \Lambda_{eu} \\ \Lambda_{ue} & \Lambda_{uu} \end{pmatrix},$$

$$\mathcal{D} = \text{diag}(\sigma_Y, \sigma_{X_1}, \sigma_{X_2}, \ldots, \sigma_{X_k}), \text{ and}$$

$$\Sigma = \begin{pmatrix} \sigma_e^2 & \Sigma_{eu} \\ \Sigma_{ue} & \Sigma_{uu} \end{pmatrix}.$$

The estimator analogous to that defined in (2.3) is

$$\tilde{\beta} = \hat{H}^{-1}(n^{-1} X'Y - D \Lambda_{ue} s_Y), \qquad (2.5)$$

where $\hat{H}$ and D are defined below equation (2.3) with Λ_{uu} replacing Λ and $G^{\dagger}$ replacing G. If λ_{ee} is unknown it is set equal to $\Lambda_{eu} \Lambda_{uu}^{-1} \Lambda_{ue}$ in the calculation of $\hat{f}$.

<u>Theorem 2</u>: Let model (2.1) with assumptions (v) and (vi) hold. Then

$$n^{\frac{1}{2}} \Sigma_{xx}(\tilde{\beta} - \beta) \xrightarrow{\mathcal{L}} N(0, A),$$

where

$$A = E\{d_t' d_t\},$$

$$d_t = (d_{t1}, d_{t2}, \ldots, d_{tk}),$$

$$d_{ti} = X_{ti} v_t - \frac{1}{2}\left[\lambda_{u_i e}\left(\frac{\sigma_Y}{\sigma_{X_i}} X_{ti}^2 + \frac{\sigma_{X_i}}{\sigma_Y} Y_t^2\right) - \sum_{j=1}^{k} \lambda_{u_i u_j} \beta_j \left(\frac{\sigma_{X_i}}{\sigma_{X_j}} X_{tj}^2 + \frac{\sigma_{X_j}}{\sigma_{X_i}} X_{ti}^2\right)\right]$$

X_{ti} is the ti^{th} element of X, v_t is the t^{th} element of v, and $\lambda_{u_i e}$ is the i^{th} element of Λ_{ue}.

The form of the result presented in Theorem 2 suggests an estimator of the variance of $\tilde{\beta}$ that is relatively easy to compute.

<u>Theorem 3</u>: Let model (2.1) with assumptions (v) and (vi) hold. Then $\hat{H}^{-1} A \hat{H}^{-1}$, where

$$\tilde{A} = (n-k)^{-1} \sum_{t=1}^{n} \tilde{d}_t' \tilde{d}_t,$$

$$\tilde{d}_t = (\tilde{d}_{t1}, \tilde{d}_{t2}, \ldots, \tilde{d}_{tk}),$$

$$\tilde{d}_{ti} = X_{ti} \hat{v}_t - \frac{1}{2}\left[\lambda_{u_i e}\left(\frac{s_Y}{s_{X_i}} X_{ti}^2\right) + \frac{s_{X_i}}{s_Y} Y_t^2 - \sum_{j=1}^{k} \lambda_{u_i u_j} \tilde{\beta}_j \left(\frac{s_{X_i}}{s_{X_j}} X_{tj}^2 + \frac{s_{X_j}}{s_{X_i}} X_{ti}^2\right)\right],$$

$$\hat{v}_t = Y_t - \sum_{j=1}^{k} \tilde{\beta}_j X_{tj},$$

is a consistent estimator of the covariance matrix of the limiting distribution of $n^{\frac{1}{2}}(\tilde{\beta} - \beta)$.

3. EXAMPLE

To illustrate the computations associated with the correction for attenuation, we use some data studied by Featherman [6] and Kelley [12]. (See also the <u>Comments</u> section of <u>The American Sociological Review</u> (1973, p. 785-796.) The data were kindly made available by Professor Featherman. The data pertain to the careers of 715 white native American urban married males. The reader is referred to the cited articles for a complete description of the data. Two of the several equations estimated in the original studies are:

$$Q_3 = \beta_1 Q_F + \beta_2 T + \beta_3 Q_1 + \beta_4 Q_2$$

$$Q_2 = \alpha_1 Q_F + \alpha_2 T + \alpha_3 Q_1 + \alpha_4 I_1$$

where

- Q_i, $i = 1,2,3$ is occupation at time i, where i = 1 is at marriage, time 2 is about eight years after marriage and time 3 is about sixteen years after marriage.
- Q_F is father's occupation
- T is years of formal education
- I_1 is income in thousands of dollars at time one.

Occupation is recorded on an eleven point scale based upon the 1947 National Opinion Research Center study [13].

Kelley gave the reliabilities for the variables as 0.718 for father's occupation, 0.933 for education, 0.861 for occupation, and 0.852 for income.

Considerable interest centered on the coefficients β_3 and α_4. Under one theoretical model both of these coefficients were hypothesized to be zero. Estimates of the two equations are given below.

$$\begin{array}{lllll} Q_3 = & 0.094Q_F + & 0.137T - & 0.044Q_1 + & 0.661Q_2 \\ & (0.040) & (0.035) & (0.074) & (0.085) \\ & (0.043) & (0.034) & (0.078) & (0.091) \end{array}$$

$$Q_2 = \underset{\substack{(0.036)\\(0.034)}}{0.080Q_F} + \underset{\substack{(0.038)\\(0.034)}}{0.176T} + \underset{\substack{(0.040)\\(0.054)}}{0.651Q_1} - \underset{\substack{(0.030)\\(0.026)}}{0.097I_1}$$

The first set of numbers in parentheses are the estimated standard errors computed under the assumption of normality. The second set are the estimated standard errors computed under the more general assumptions. The coefficients are reported in the original units. Also the method used to treat missing values differed from that used by Featherman. Therefore, the coefficients are not identical to those reported by Featherman and Kelley. From a substantive viewpoint the coefficient for Q_1 in equation one could easily be zero. However, if one accepts the assumptions it is very unlikely that the coefficient for income in the second equation is zero.

The variables are clearly not normal because all are restricted to a few integer values. Procedures based on normality gave estimated standard errors very similar to those obtained under the more general assumptions for the first equation. On the other hand, the estimated standard error for Q_1 in the second equation computed under the normal assumption is quite different from that computed under the more general assumptions.

The joint distribution of Q_2 and Q_1 deviates considerably from normality. For example, the residuals from the ordinary regression of Q_2 on Q_1, say $\hat{\delta}$, have a coefficient of skewness of 0.34 and a kurtosis of 3.32. The approximate standard errors of these quantities, under normality, are 0.09 and 0.18, respectively. There is also considerable evidence that the conditional mean of Q_2 given Q_1 is not linear, the t statistic for the quadratic term in a regression of Q_2 on Q_1 and Q_1^2 being 6.7. Also the conditional variance of Q_2 given Q_1 is not constant, the regression of the squared regression residuals, $\hat{\delta}^2$, on Q_1 and Q_1^2 give an F-statistic of 30.8 with two and 712 degrees of freedom.

Because of the robustness of the general procedure it is recommended unless the sample size is very small.

ACKNOWLEDGEMENT

This research was partly supported by the Bureau of the Census through Joint Statistical Agreements J.S.A. 75-1, and J.S.A. 76-66.

REFERENCES

[1] Blalock, H. M., Jr., "Estimating Measurement Error Using Multiple Indicators and Several Points in Time," American Sociological Review, 35 (February 1969), 101-11.

[2] Bock, R. D. and Peterson, A. C., "A Multivariate Correction for Attenuation," Biometrika, 62 (December 1975), 673-78.

[3] Bohrnstedt, G. W. and Carter, T. M., "Robustness in Regression Analysis," in H. L. Costner, ed., Sociological Methodology, 1971, San Francisco: Jossey-Bass, Inc., 1971, 118-46.

[4] Booth, G. D., "The Errors-in-Variables Model When the Covariance Matrix is Not Constant," Ph.D. dissertation, Iowa State University, 1973.

[5] Cochran, W. G., "Errors of Measurement in Statistics," Technometrics, 10 (November 1968), 637-66.

[6] Featherman, D. L., "A Research Note: A Social Structural Model for the Socioeconomic Career," Amer. J. Soc., 77 (1971), 293-304.

[7] Fuller, W. A., "Properties of Estimators in the Errors-in-Variables Model," paper presented at the 1971 annual meeting of the Econometric Society, 1971 (mimeographed).

[8] Hidiroglou, M. A., "Estimation of Regression Parameters for Finite Populations," Ph.D. dissertation, Iowa State University, 1974.

[9] Johnston, J., Econometric Methods, New York: McGraw-Hill Book Co., 1963.

[10] Jöreskog, K. G., "A General Method for Analysis of Covariance Structures," Biometrika, 57 (August 1970), 239-51.

[11] Jöreskog, K. G., "Statistical Analysis of Sets of Congeneric Tests," Psychometrika, (June 1971), 109-33.

[12] Kelley, J., "Causal Chain Models for the Socioeconomic Career," American Sociological Review, 38 (1973), 481-93.

[13] National Opinion Research Center, "Jobs and Occupations: A Popular Evaluation," Opinion News, 9 (September 1947), 3-13.

[14] Malinvaud, E., Statistical Methods of Econometrics, Amsterdam: North-Holland Publishing Co., 1966.

[15] Rao, C. R., Linear Statistical Inference and Its Applications, New York: Wiley, 1965.

[16] Walker, H. M. and Lev, J., Statistical Inference, New York: Holt, Rhinehart & Winston, Inc., 1953.

[17] Warren, R. D., Keller, J. K., and Fuller, W. A., "An Errors-in-Variables Analysis of Managerial Role Performance," J. Amer. Statist. Assoc., 69 (December 1974), 886-93.

[18] Wiley, D. E., "The Identification Problem for Structural Equations Models with Unmeasured Variables," in A. S. Goldberger and O. D. Duncan, eds., Structural Equation Models in the Social Sciences, New York: Seminar Press, 1973.