THE EFFECT OF MEASUREMENT ERROR ON REGRESSION DIAGNOSTICS

Marian K. Snyder
University Undergraduate Fellow, 1986-1987
Texas A&M University
Department of Statistics

APPROVED:

Fellows Advisor Fred Dahm

Honors Director Lawrence Clayfield

THE EFFECT OF MEASUREMENT ERROR ON REGRESSION DIAGNOSTICS

ABSTRACT

Regression diagnostics detect outliers. The purpose of this study is to analyze the effect of measurement error on the detection process of the regression diagnostic, the hat diagonals. The hat diagonals are analyzed according to four cases: the error-free and contamination-free case (Model 1), the measurement error case (Model 2), the outlier case (Model 3), and the combined measurement error and outlier case (Model 4). The analysis leads to the limiting probability distribution of $(n-1)h_{ii}$ under Model 4. This distribution is simulated for sample sizes, N, equal to 20, 40, and 60.

Introduction

Regression is a widely used statistical tool used to represent an assumed linear relationship between a response variable, Y, and an explanatory variable, x. Least squares methodology is employed to estimate the linear relationship between Y and x based upon a sample (Y_i, x_i), i=1, ..., n. The simplicity and ease of using regression packages allows insidious error to creep into the analysis. Many researchers are unfamiliar with the assumptions and limitations of regression. Therefore, we propose to study one tool that allows for detection of the invalidity of least squares: the regression diagnostic.

Regression diagnostics and outliers

According to Hocking(1983), regression diagnostics are numerical functions of the data that may be used to detect inputs, x_i, that are "far from the bulk of the data" (Hocking, p. 222). The points are collectively termed outliers. A single outlier has the capability of invalidating the line of best fit. Therefore, outliers are sometimes referred to as high leverage points because of this capability. Diagnostics can be used to detect any invalidation of least squares due to outliers.

Diagnostics for the single X case are trivial. Diagnostics are extremely useful for the p-dimensional case when plots become cumbersome in detecting outliers. Therefore, the study of regression diagnostics is important for the more involved case where the X matrix contains p columns.

Measurement error

Measurement error can also invalidate the least squares fit of the line. On average, data with measurement error will underestimate the slope of the straight-line relationship between Y and x if least squares methodology is used. Human measurement error such as misrecording data, using subjective measurements, or using inaccurate machinery are common occurrences. For these reasons measurement error must be considered in the regression analysis.

Specifically, we wish to analyze the effect of measurement error on regression diagnostics. We expect to see that measurement error masks the detection of outliers. As a result of the spreading out of the data due to measurement error, the outlier would become more a part of the bulk of the data. Thus, its distance from the line of best fit would no longer be significant.

The hat diagonals

We will focus our attention upon one diagnostic, the hat diagonals. The i^{th} diagonal of the hat matrix, H, is denoted by h_{ii}. The matrix H is an nxn projection matrix where $H = X(X'X)^{-1}X'$. According to Dunn (1982), the "hat matrix provides a direct measure of extreme observations" (Dunn, p. 17). He explains that when an observation, x_i, is extreme, the consequent h_{ii} is large in response to the exaggerated distance of of the x_i from the bulk of the data. Observations, x_i, that result

in large h_{ii} values are detected as high leverage points (Dunn, p. 17). Our interest in the hat diagonals is in its performance under the stress of outlier contamination and measurement error.

The Literature Review

Draper and Smith (1966) offer an introductory and complete explanation of regression. Hocking (1983) provides an overview of the developments in linear regression over the past twenty years. In particular, Hocking discusses regression diagnostics. He points out that the hat diagonal is a widely used diagnostic for detecting outliers in the explanatory variables. Dunn (1982) gives detailed discussions of several diagnostics. However, there appears to be no reference to the affect of measurement error on the regression diagnostics. Fuller (1981) provides a comprehensive treatment of measurement error models.

The Analysis

The purpose of this section is to examine the limiting distribution of the hat diagonals under the influence of outliers and measurement error. To accomplish this, we analyze the hat diagonals for the following four models: the error-free and contamination-free case (Model 1); the measurement error case (Model 2); the outlier case (Model 3); and, finally, the combined measurement error and outlier case (Model 4).

In order to begin the analysis, we must define the concept of convergence in probability. Serfling (1980) defines convergence in probability as follows.

Definition The random variable Y_n converges in probability to the random variable Y if and only if, for every $\varepsilon > 0$, the $\lim_{n\to\infty} P(|Y_n - Y| < \varepsilon) = 1$. The techniques used for demonstrating convergence in probability are derived in Chapter 1 of Serfling (1980).

The Models

In order to analyze the hat diagonals for each case in the analysis, we must define four models to coincide with each case.

Model 1: for the error-free and contamination free case.

Let

$$y_i = \beta_0 + \beta_1 x_i$$

$$Y_i = y_i + e_i \qquad i = 1, 2, \ldots, n$$

$$(x_i, e_i) \text{ iid } NI((\mu_x, 0)', \text{diag } (\sigma_{xx}, \sigma_{ee}))$$

This is the simple regression model where x_i is the true, observed variable.

Model 2: for the measurement error case

Let

$$y_i = \beta_0 + \beta_1 x_i$$

$$Y_i = y_i + e_i$$

$$X_i = x_i + u_i$$

$$(x_i, e_i, u_i) \text{ iid } NI((\mu_x, 0, 0)', \text{diag } (\sigma_{xx}, \sigma_{ee}, \sigma_{uu}))$$

where x_i is the observed independent variable and u_i is the is the measurement error.

Model 3: for the outlier case

Let

$$y_i = \beta_0 + \beta_1 x_i$$

$$Y_i = y_i + e_i \qquad i = 1, 2, \ldots, n$$

where e_i iid $N(0, \sigma_{ee})$

and

$$x_i \sim (1 - \delta_x)\, N(\mu_x, \sigma_{xx}) + \delta_x\, N(\mu_x, \kappa^2_x \sigma_{xx})$$

$$0 > \delta_x \geq 1, \quad \sigma_{xx} > 0, \quad \kappa_x^2 >> 1$$

The model allows that with probability δ_x, the i^{th} realization of x_i will be drawn from the contaminating $N(\mu_x, \kappa^2_x \sigma_{xx})$ distribution.

Model 4: for the combined measurement error and outlier case.

Let

$$y_i = \beta_0 + \beta_1 x_i$$

$$Y_i = y_i + e_i$$

$$X_i = x_i + u_i \qquad i = 1, 2, \ldots, n$$

where

$$x_i \sim (1 - \delta_x)\, N(\mu_x, \sigma_{xx}) + \delta_x\, N(\mu_x, \kappa_x^2\, \sigma_{xx})$$

$$0 \leq \delta_x \leq 1, \quad \kappa_x^2 >> 1, \quad \sigma_{xx} > 0,$$

e_i iid $N(0, \sigma_{ee})$, $\sigma_{ee} > 0$,

and

u_i iid N (0, σ_{uu}) , $\sigma_{uu} \geq 0$,

where x_i is chosen from the N(μ_x, κ_x^2 σ_{xx}) with probability δ_x and x_i is the observed value. Then, u_i remains as the measurement error variable.

The Cases

The Error-Free and Contamination-Free Case.

This first case studies the hat diagonals using Model 1. Given that the i^{th} diagonal element of the hat matrix H, h_{ii}, can be expressed as

$$(n-1)\ h_{ii} = \frac{(\bar{X}_i - \bar{X})^2}{(n-1)^{-1} \sum_{j=1}^{n} (X_j - \bar{X})^2}$$

for the single X case, it is instructive to look at the separate limiting distributions of the numerator and denominator. We must, then, examine the expected values of each. We obtain that

$$E[(X_i - \bar{X})^2] = (\frac{n-1}{n})\ \sigma_{xx}$$

since X is iid N (μ_x, σ_{xx}) and $E[\sum_{j=1}^{n}(X_j - \bar{X})^2] = (n-1)\sigma_{xx}$. From this we conclude that $(X_i - \bar{X})^2$ converges in probability to the square of a N(0, σ_{xx}) random variable. In addition, $(n-1)^{-1} \sum_{j=1}^{n}(X_j - \bar{X})^2$ converges in probability to σ_{xx}. Thus, $(n-1)\ h_{ii}$ converges in probability to a chi-square random

variable with one degree of freedom. If the X matrix has column dimension, p, $(n-1)\ h_{ii}$ can be shown to converge in probability to a chi-square random variable with p degrees of freedom. These results are derived in Appendix A.

The Measurement Error Case

The analysis of the hat diagonals in the presence of measurement error, as specified in Model 2, is essentially the same as that of Case 1. The only change occurs in the model as found in X = x + u where x is the true observed value and u is the measurement error. The variance - covariance matrix of the observed independent variable X is Σ_{XX} where

$$\Sigma_{XX} = \Sigma_{xx} + \Sigma_{uu} \quad .$$

However, the arguments leading to the result in Case 1 are the same, with Σ_{XX} replacing Σ_{xx}. Hence, $(n-1)\ h_{ii}$ converges in probability to a chi-square random variable with p degrees of freedom as in Case 1. See Appendix B for details.

The Outlier Case

This case involves a more indepth analysis than the preceding two cases. We must now draw upon Model 3 where

$$x_i \sim (1-\delta_x)\ N\ (\mu_x,\ \sigma_{xx}) + \delta_x\ N(\mu_x,\ \kappa_x^2\ \sigma_{xx})$$

$$0 < \delta_x \leq 1,\ \sigma_{xx} > 0,\ \kappa_x^2 >> 1$$

The model allows that with probability δ_x, the i^{th}

observation will be drawn from the contaminating $N(\mu_x, \kappa_x^2 \sigma_{xx})$ distribution. For the single X case the numerator of $(n-1)h_{ii}$ converges in probability to a mixture of a $N(0, \sigma_{xx})$ random variable and a $N(0, \kappa_x^2 \sigma_{xx})$ random variable The denominator of $(n-1)h_{ii}$ converges to a constant, as shown in Appendix C.

The Combined Case

We must now consider the combined case of measurement error along with outliers as in Model 4. The notation changes from Case 3 include

$$X_i = x_i + u_i$$

where

$$u_i \sim \text{iid } N(0, \sigma_{uu}) \text{ random variable}$$

The conclusions of this analysis concur with the outlier case with slight alteration. That is, $X_i - \bar{X}$ converges in probability to a mixture of the $N(0, \sigma_{xx} + \sigma_{uu})$ and $N(0, \kappa^2\sigma_{xx} + \sigma_{uu})$ random variables as shown in Appendix D.

The Simulation

The purpose of the simulation is to examine the small sample behavior of $(n-1)h_{ii}$. We generate the observed variable, x_i, and the measurement error variable, u_i, according to Model 4. The $X_i = x_i + u_i$ are used to compute the h_{ii}. We then arbitrarily select h_{11} to be stored. After storing five thousand independent

observations of (n-1) h_{11} we computed the empirical .01, .05, .10, .90, .95, and .99 percentiles of the (n-1) h_{ii}'s.

Results and Discussion

The limiting distribution of the (n-1) h_{ii}

From the analysis in Appendix E we are able to determine the limiting probability distribution of (n-1) h_{ii}. Recall that

$$(n-1)\, h_{ii} = \frac{(X_i - \bar{X})^2}{(n-1)^{-1} \sum_{j=1}^{n} (X_j - \bar{X})^2} .$$

We have that the denominator converges in probability to a constant, q, where

$$q = \sigma_{xx}(1 - (\kappa_x^2 - 1)\delta_x) + \sigma_{uu}.$$

The numerator converges to the square of a mixture of normal random variables that the limiting distribution function of (n-1) h_{ii} can be expressed as

$$F(\eta) = \lim_{n\to\infty} P[(n-1)h_{ii} \leq \eta] = P(-\sqrt{q\eta} \leq X_i - \mu \leq \sqrt{q\eta})$$

$$= P(X_i - \mu \leq \sqrt{q\eta}) - P(X_i - \mu < -\sqrt{q\eta})$$

Table 2 is a table of $F_n(\eta)$ for the case where

$$\sigma_{xx} = 1, \quad \sigma_{uu} = 1, \quad \delta_x = .05, \quad \mu_x = 0, \text{ and } \kappa_x^2 = 10 .$$

The .01, .05, and .10 values of $F(\eta)$ occur at η = 0.00022, 0.00538, and 0.02160, respectively. The η values for $F(\eta)$ = .90, .95, and .99 are 3.92300, 5.83500, and 13.16000, respectively.

When $\delta_x = 0$, the limiting distribution of $(n-1)h_{ii}$ is chi-square with one degree of freedom. This follows from the fact that Model 4 reduces to Model 2 when $\delta_x = 0$. Table 1, serves two purposes. First, it demonstrates that as n gets larger, the empirical percentiles generally tend toward the percentiles of the limiting distribution. Second, because the empirical percentiles are based on finite samples, sampling variability may cause the empirical percentiles to deviate slightly from the expected trend. For example, the empirical .95 percentiles for N=20, 40, 60 are, respectively, 3.56, 3.80, 3.69. Apparently sampling variability causes the N=60 empirical .95 percentile to be farther from the limiting .95 percentile (3.84) than the N=40 empirical .95 percentile.

We know that for models 1 and 2 that $(n-1)\ h_{ii}$ tends toward a chi-square random variable with one degree of freedom. We expect the combined case of outliers and measurement error to deviate from the chi-square. Table 2 lists the results of the analysis and the simulation of the limiting distribution of $(n-1)\ h_{ii}$ when $\delta_x = 0$ for Model 4. The first column of the table shows the true values of the limiting distribution while the other three columns show empirical percentiles computed by simulation.

TABLE 1. The limiting distribution of $(n-1)h_{ii}$ with $\delta_x = 0$

α	Limiting Distribution α Percentile	Simulation Percentage point N=20	N=40	N=60
0.01	0.000157	0.0001961	0.0002394	0.0001278
0.05	0.003934	0.0038784	0.0040793	0.0042236
0.10	0.015790	0.0151803	0.0161976	0.0042236
0.90	2.705600	2.542054	2.636511	2.6817961
0.95	3.841500	3.5596993	3.8057005	3.6894607
0.99	6.635000	5.5898714	5.8526012	5.9616838

Note: Parameter values are: $\delta_x = 0$

$\sigma_{xx} = 1$

$\sigma_{uu} = 1$

$\mu_x = 0$

$\kappa_x^2 = 10$

Table 2. The limmiting Distribution Results of $(n-1)h_{ii}$

α	Limiting Distribution α Percentile	Simulated Distributions N=20	N=40	N=60
0.01	0.00022	0.00014	0.00012	0.00013
0.05	0.00538	0.00299	0.00270	0.00298
0.10	0.02160	0.01102	0.01153	0.01196
0.90	3.92300	2.227256	2.50410	2.32620
0.95	5.83500	3.20180	4.41760	3.71620
0.99	13.16000	5.08360	15.23107	17.96290

Note: Parameter values are $\delta_x = 0.05$,

$\sigma_{xx} = 1$

$\sigma_{uu} = 1$

$\mu_x = 0$

$\kappa_x^2 = 10$

The Simulation Results.

Table 2 lists results of the analysis and the simulation of the limiting distribution of $(n-1)h_{ii}$ when $\delta_x = .05$. Columns in Table 2 are defined the same as those in Table 1. The limiting percentiles in Table 2 differ substantially from those of Table 1. The effect of 5% ($\delta_x = .05$) contamination by a N(0,10) population of true x values causes the limiting distribution to become skewed to the right. Comparison of the limiting percentiles to the empirical percentiles within Table 2 also reveals that empirical quantiles do not converge (as sample size increases) nearly as fast when contamination occurs. For example, even when N=60 in Table 2, empirical .90, .95, and .99 quantiles are 2.33, 3.71, and 17.96, respectively. The corresponding limiting quantiles are 3.92, 5.84, and 13.16, respectively.

Another aspect of our study is the effect of measurement errors on regression diagnostics. Unfortunately, time constraints have not allowed us to analyze adequately this aspect and results are not presented here.

Conclusion

We have proposed to study the effect of measurement error on regression diagnostics. We have partially fulfilled that commitment. We have defined the probability distribution of the hat diagonals in the presence of random measurement error and on the limiting distribution of the h_{ii} under the same conditions. Due to a time constraint, however, we were unable to complete the

analysis of the hat diagonals with measurement error variation. Thus, the effect of the measurement error on the hat diagonals is incomplete.

References

Anderson, T.W. (1958), An Introduction to Multivariate Statistical Analysis, New York: John Wiley.

Dunn, Mark R. (1982), Regression Diagnostics. Unpublished Ph.D. dissertation. Texas A&M University, College Station, Texas.

Draper, N.R., and Smith, H. (1981), Applied Regression Analysis (2nd Ed.), New York: John Wiley.

Fuller, W.A. (1981), Measurement Error Models. Forthcoming to be published by John Wiley.

Hocking, R.R. (1983), "Developments in Linear Regression Methodology: 1959-1982," Technometrics, 25, 219-230.

Serfling, R.J. (1980), Approximation Theorems of Mathematical Statistics, New York: John Wiley.

APPENDIX A

In this appendix we derive an expression for, and properties of, $(n-1)h_{ii}$, where h_{ii} is the i^{th} diagonal of the centered hat matrix H_c. First we consider the single explanatory variable case in detail. Then we sketch the derivations for multiple explanatory variables.

The simple linear regression model (Model 1) is

$$Y_i = \beta_0 + \beta_1 x_1 + e_i \ , \quad i=1,2,\ldots,n,$$

where we assume the x_i's are iid $N(\mu_x, \sigma_{xx})$ random variables. Written in deviation form the model becomes

$$(Y_i-\bar{Y}) = \beta_1(x_i-\bar{x}) + e_i \ , \ i=1,2,\ldots,n,$$

or, in obvious matrix notation,

$$Y = \beta_1 x_c + \varepsilon \ .$$

The centered hat matrix for the single explanatory variable model is defined to be

$$H_c = x_c(x_c'x_c)^{-1}x_c' \ ,$$

with i^{th} diagonal element

$$h_{ii} = (x_i - \bar{x})^2/\Sigma_{j=1}^{n} (x_i-\bar{x})^2.$$

Therefore

$$(n-1)h_{ii} = (x_i-\bar{x})^2/[(n-1)^{-1}\Sigma_{j=1}^{n}(x_i-\bar{x})^2] \quad . \qquad (1)$$

Next, we consider large sample properties of $(n-1)h_{ii}$ for the single-x case. The denominator of (1) converges in probability to σ_{xx}. The numerator of (1) converges in probability to $(x_i-\mu_x)^2$, i.e., the square of a $N(0,\sigma_{xx})$ random variable. Therefore $(n-1)h_{ii}$ converges in probability to the square of a N(0,1) random variable. Thus, the limiting distribution of $(n-1)\ h_{ii}$ is chi-square with one degree of freedom.

If κ explanatory variables appear in the regression model, the model in derivation form is

$$(Y_i - \bar{Y}) = \beta_1(x_{i1}-\bar{x}_1) +\ldots+\beta_\kappa(x_{i\kappa}-\bar{x}_\kappa) + \varepsilon_i,\ i=1,2,\ldots,n \ .$$

In matrix notation, the model in deviation form, is

$$Y_c = x_c\beta_c + \varepsilon \ .$$

The centered hat matrix for the κ-variable case is $H_c=x_c(x_cx_c)^{-1}x_c'$. Although it is infeasible to derive an explicit expression for $(n-1)h_{ii}$, the i^{th} diagonal element of H_c, the limiting distribution of $(n-1)h_{ii}$ can still be derived. First consider

$$S_{xx} = (n-1)^{-1}(x_c'x_c)$$

$$= \text{sample covariance matrix of the vectors } (x_{i1},\ldots,x_{i\kappa})'$$

Clearly, S_{xx} converges in probability to Σ_{xx}, the covariance matrix of $(x_{i1},\ldots,x_{i\kappa})'$. Also

$$[(x_{i1}-\bar{x}_1),\ldots,(x_{i\kappa}-\bar{x}_\kappa)]' \xrightarrow{P} [(x_{i1}-\mu_1),\ldots,(x_{i\kappa}-\mu_\kappa)]',$$

where μ_j is the mean of x_{ij}, $j=1,2,\ldots,\kappa$. Therefore,

$$(n-1)h_{ii} \xrightarrow{P} k[(x_{i1}-\mu_1),\ldots,(x_{i\kappa}-\mu_\kappa)]\ \Sigma_{xx}^{-1}[(x_{i1}-\mu_1),\ldots,(x_{1\kappa}-\mu_\kappa)]'$$

$$= Z_i'Z_i$$

where Z_i is a vector of $N_\kappa(0,I)$ random variables. Thus,

$$(n-1)h_{ii} \xrightarrow{P} \chi^2_\kappa .$$

APPENDIX B

If Model 2 is the true underlying model for the data, the arguments of Appendix A carry over trivially to this case as well. Let $\Sigma_{XX} = \Sigma_{xx} + \Sigma_{uu}$, where Σ_{uu} is the covariance matrix of the vector of measurement errors. Then Σ_{XX} is the covariance matrix of observed explanatory variables, because $X = x + u$, where x and u are independent normal vectors. Therefore, for the centered hat matrix $H_c = X_c(X_c'X_c)X_c'$ of the observed ($X = x + u$) explanatory variables,

$$S_{XX} = (n-1)^{-1}X_c'X_c$$

converges in probability to Σ_{XX}. Also, the i^{th} row of X_c converges in probability to a $N(0,\Sigma_{XX})$ random vector. Thus, $(n-1)h_{ii}$ converges in probability to a χ^2_κ random variable

APPENDIX C

Assume Model 3 holds. This is the outlier model. To consider properties of $(n-1)\, h_{ii}$ for this case define the following notation.

Let $W \sim N(\mu_x, \sigma_{xx})$ and $Z \sim N(\mu_x, \kappa_x^2 \sigma_{xx})$, W and Z independent. Define V, independent of W, Z to be

$$V = \begin{cases} 1 \text{ with probability } 1-\delta_x \\ 0 \text{ with probability } \delta_x \end{cases} .$$

where $0 < \delta_x \leq 1$

Then

$$X = VW + (1-V)Z$$

for our model and

$$\begin{aligned} E(X) &= E_V E(X|V = v) \\ &= (1-\delta_x)E(X|V=1) + \delta_x E(X|V=0) \\ &= (1-\delta_x)\mu + \delta_\mu \\ &= \mu \ . \end{aligned}$$

Also,

$$\begin{aligned} E(X^2) &= E_V E(X^2|V = v) \\ &= (1-\delta_x)E(X^2|V = 1) + \delta_x E(X^2|V = 0) \\ &= (1-\delta_x)E(W^2) + \delta_x E(Z^2) \\ &= (1-\delta_x)(\sigma_{xx} + \mu_x^2) + \delta_x(\kappa_x^2\sigma_{xx} + \mu_x^2) \\ &= \sigma_{xx}(1 + (\kappa_x^2 - 1)\delta_x) + \mu_x^2 \end{aligned}$$

Therefore, because

$$\bar{X}^2 \xrightarrow{p} [\ E(X)\ ]^2$$

and

$$(n-1)^{-1} \sum_{j=1}^{n} X_j^2 \xrightarrow{p} E(X^2)\ ,$$

we have

$$(n-1)^{-1} \left[\sum_{j=1}^{n} (X_j - \bar{X})^2\right] = (n-1)^{-1}\left[\sum_{j=1}^{n} X_j^2 - n\,\bar{X}^2\right]$$

$$\xrightarrow{p} E(X^2) - [\ E(X)\ ]^2$$

$$\xrightarrow{p} \sigma_{xx}(1 - (k_x^2 - 1)\delta_x).$$

Also,

$$X_i - \bar{X} \xrightarrow{p} X_i - E(X_i) = X - \mu.$$

Now,

$$\begin{aligned} P(X - \mu < c) &= P(V = 1) \cdot P(X - \mu_x < c \mid V = 1) \\ &\quad + P(V = 0) \cdot P(X - \mu_x < \mid V = 0) \\ &= (1 - \delta_x)\ P\ (W - \mu < c) + \\ &\quad + \delta_x \cdot P(Z - \mu_x < c) \\ &= (1 - \delta_x) \cdot P\ [(W - \mu_x)/\sigma_x < c/\sigma_x] \\ &\quad + \delta_x \cdot P\ [(z - \mu_x)/\sigma_x < c/\sigma_x] \\ &= (1 - \delta_x)\Phi\ (c/\sigma_x) + \delta_x\ \Phi\ (c/\kappa_x\sigma_x) \end{aligned}$$

That is, $X_i - \bar{X}$ converges in probability to a mixture of a $N(0, \sigma_{xx})$ random variable and a $N(0, \kappa_X^2 \sigma_{xx})$ random variable.

APPENDIX D

Assume Model 4 holds. Then write

$$X_i = x_i + u_i$$

and

$$x_i = V_i W_i + (1 - V_i) Z_i$$

where V_i's are iid random variables with p.d.f.

$$V_i = \begin{cases} 1 \text{ with probability } 1 - \delta_x \\ 0 \text{ with probability } \delta_x \end{cases},$$

W_i's are iid random variables from N (μ_x, σ_{xx}) population,

Z_i's are iid random variables from N (μ_x, $\kappa_x^2 \sigma_{xx}$) population,

and

u_i's are iid random variables from a N(0, σ_{uu}) population.

Then,

$$E(X_i) = E(x_i + u_i)$$

$$= E_{V_i} E(x_i | V = v) + E(u_i)$$

$$= \mu_x + 0$$

$$= \mu_x$$

from Appendix C. Further,

$$E(X_i^2) = E_{V_i} E(X_i^2 | V = v)$$

$$= E_{V_i} E[(X_i + u_i)^2 | V = v]$$

$$= E_{V_i} E(x_i^2|V=v) + 2E_{V_t} E(u_i)E(x_i)$$

$$+ E_{V_i} E(u_i^2|V=v)$$

Since $E(u_i) = 0$,

$$E(X_i^2) = \sigma_{xx}(1 + (k_x^2 - 1)\delta_x + \mu_x^2 + \sigma_{uu}.$$

Therefore, because

$$\bar{X}^2 \xrightarrow{p} [E(X)]^2$$

and

$$(n-1)^{-1} \Sigma_{j=1}^{n} X_j^2 \xrightarrow{p} E(X^2)$$

we have that

$$(n-1)^{-1} \Sigma_{j=1}^{n} X_j^2 - n\bar{X}^2] \xrightarrow{p} E(X^2) - [E(X)]^2,$$

as before.

Comparing these results with those of Appendix C, we see that the only change is the greater variance due to the measurement error. That is, $X_i - \bar{X}$ converges in probability to a mixture of a $N(0, \sigma_{xx} + \sigma_{uu})$ random variable and a $N(0, \kappa_x^2\sigma_{xx} + \sigma_{uu})$ random variable.

APPENDIX E

According to model 4, the distribution of $(n-1)h_{ii}$ can be derived.

$$P[(n-1)h_{ii} \leq \eta]$$

$$= [X_i - \mu]^2/c \leq \eta]$$

where $c = \sigma_{xx}(1-(k_x^2 - 1)\delta_x) + \sigma_{uu}$

Then,

$$P[(n-1)h_{ii} \leq \eta] = P[0 \leq (X_i - \mu_x)^2 \leq q\eta]$$

$$= P[-\sqrt{c\eta} \leq (X-\mu_x) \leq \sqrt{q\eta}]$$

$$= P[(X-\mu_x) \leq \sqrt{q\eta}]$$

$$- P[)X-\mu_x) \leq - \sqrt{q\eta}].$$

The separate probabilities then are calculated by

$$P[(x-\mu_x) < L] = (1-\delta_x)\Phi[L/\kappa_{xx}^2\sigma_{xx} + \sigma_{uu})^{1/2}]$$

$$+ \delta_x\Phi[-L/(\kappa_x^2\sigma_{xx} + \sigma_{uu})^{1/2}$$

where L is $\sqrt{q\eta}$.