I. Bias due to Errors in Variables

Suppose Population relationship is:

$$(1) \qquad Y_i = \alpha + \beta X_i + \varepsilon_i$$

Our measures of Y and X contains errors so the observed values are:

$$(2) \qquad Y_i^\star = Y_i + v_i$$

$$(3) \qquad X_i^\star = X_i + w_i$$

where $E(v_i, v_j) = 0, E(w_i, w_j) = 0, E(v_i, w_j) = 0, E(v_i, \varepsilon_i) = 0, E(w_i, \varepsilon_j) = 0$
So the measurement errors are truly random noting

$$(4) \quad Y_i = Y_i^\star - v_i$$

$$(5) \quad X_i = X_i^\star - w_i$$

Substituting (4) and (5) into (1), we get

$$(6) \quad Y_i^\star - v_i = \alpha + \beta(x_i^\star - w_i) + \varepsilon_i$$

so

$$(7) \quad Y_i^\star = \alpha + \beta x_i^\star + \varepsilon_i^\star$$

where

$$\varepsilon_i^\star = \varepsilon_i + v_i - \beta w_i$$

Thus if we do ordinary least squares we will estimate (7). We evaluate covariance of $x^\star$ and $\varepsilon^\star$ and find

$$(9) \quad Cov(x_i^\star, \varepsilon_i^\star) = E[x_i^\star - E(x_i^\star)]\varepsilon_i^\star = E[w_i(\varepsilon_i^\star + v_i - \beta w_i)] = -\beta\sigma_{w_i}^2 \neq 0$$

The estimate of (7) will be

$$(10) \qquad Y_i = a + bX_i + e_i$$

$$(11) \qquad E(b) = \beta + E\left[\frac{\sum X_i \varepsilon_i}{\sum X_i^2} \right]$$

II. Errors in Variables

From (9) we know E$(\sum X_i \varepsilon_i) = -\beta\sigma_{w_i}^2$ so

$$(12) \quad E(b) = \beta - \beta\frac{\sigma_{w_i}^2}{\sum X_i^2}$$

therefore the least squares estimate b is biased toward zero. Note if the error in measurement is isolated only in the independent variable Y, then equation (8) becomes

$$(13) \qquad \varepsilon_i^\star = \varepsilon_i + v_i$$

and (9) becomes

$$(14) \quad Cov(x_i^\star, \varepsilon_i^\star) = E[w_i(\varepsilon_i^\star + v_i)] = 0$$

so

$$E(b) = \beta + E\left[\frac{\sum X_i \varepsilon_i}{\sum X_i^2} \right] = 0$$

and b is an unbiased estimate of b

But still

$$(15) \qquad \sigma_{\varepsilon^\star}^2 = \sigma_{\varepsilon_i}^2 + \sigma_{v_i}^2$$

So the estimated error variance is greater than the true error variance which reduces the efficiency of all the estimators.

III.Population Model in deviation notation

$y_i = \beta X_i + \varepsilon_i$

Ordinary least squares estimates: $y_i = bx_i + \varepsilon_i$

$$b = \frac{\sum y_i x_i}{\sum x_i^2}$$

substitute population value for y_i

$$(1) \qquad b = \frac{\sum x_i(\beta X_i + \varepsilon_i)}{\sum x_i^2} = \beta \frac{\sum x_i^2}{\sum x_i^2} + \frac{\sum x_i \varepsilon_i}{\sum x_i^2} = \beta + \frac{\sum x_i \varepsilon_i}{\sum x_i^2}$$

expected value of O.L.S estimate

$$(2) \qquad E(b) = E\left(\beta + \frac{\sum x_i \varepsilon_i}{\sum x_i^2}\right) = \beta + E(\frac{\sum x_i \varepsilon_i}{\sum x_i^2})$$

if X_i and ε_i are independent $E(\frac{\sum x_i \varepsilon_i}{\sum x_i^2}) = E\left(\frac{x_i \varepsilon_i}{\sum x_i^2}\right) = E\left(\frac{x_i}{\sum x_i}\right)E(\varepsilon_i) = 0$

$$(3) \qquad E(b) = \beta$$

Variance of least-squares estimate

$$(4) \qquad \mathrm{Var}(b) = E[b - E(b)]^2$$

substituting (1) and assuming (3) holds

$$(5) \quad \mathrm{Var}(b) = E\left[\beta + \frac{\sum x_i \varepsilon_i}{\sum x_i^2} - \beta\right]^2 = E\left[\frac{\sum x_i \varepsilon_i}{\sum x_i^2}\right]^2 = E\left[\frac{\sum x_i^2 u_i^2}{\sum (x_i^2)^2} + 2\frac{\sum_{i<j} x_i x_j u_i u_j}{\sum (x_i^2)^2}\right]$$

We take a diversion to illustrate (5) in the case of 3 observations

$$(5a) \quad E\left[\frac{\sum x_i \varepsilon_i}{\sum x_i^2}\right]^2 = E\left[\frac{(x_1\varepsilon_1 + x_2\varepsilon_2 + x_3\varepsilon_3)(x_1\varepsilon_1 + x_2\varepsilon_2 + x_3\varepsilon_3)}{(\sum x_i^2)^2}\right]$$

$$= E\left[\frac{x_1^2\varepsilon_1^2 + x_1x_2\varepsilon_1\varepsilon_2 + x_1x_3\varepsilon_1\varepsilon_3 + x_2x_1\varepsilon_2\varepsilon_1 + x_2x_3\varepsilon_2\varepsilon_3 + x_2^2\varepsilon_2^2 + x_3x_1\varepsilon_3\varepsilon_1 + x_3x_2\varepsilon_3\varepsilon_2 + x_3^2\varepsilon_3^2}{(\sum x_i^2)^2}\right]$$

We can summarize the six cross product terms as follows

$$(5b) \quad \begin{aligned} x_1x_2\varepsilon_1\varepsilon_2 + x_2x_1\varepsilon_2\varepsilon_1 &= 2x_1x_2\varepsilon_1\varepsilon_2 \\ x_2x_3\varepsilon_2\varepsilon_3 + x_3x_2\varepsilon_3\varepsilon_2 &= 2x_2x_3\varepsilon_2\varepsilon_3 \\ x_1x_3\varepsilon_1\varepsilon_3 + x_3x_1\varepsilon_3\varepsilon_1 &= 2x_1x_3\varepsilon_1\varepsilon_3 \end{aligned}$$

$$= 2\sum_{i<j} x_i x_j \varepsilon_i \varepsilon_j$$

$$(5c) \quad E\left[\frac{\sum x_i\varepsilon_i}{\sum x_i^2}\right]^2 = E\left[\frac{\sum x_i^2\varepsilon_i^2}{(\sum x_i^2)^2} + 2\sum_{i<j} x_i x_j \varepsilon_i \varepsilon_j\right]$$

returning to (5), assuming x_i and ε_i independent

$$(6) \; E\left[\frac{\sum_{i=1} x_i^2\varepsilon_i^2}{\sum (x_i^2)^2}\right] = E\left[\frac{\sum (x_i^2)}{\sum (x_i^2)^2}\right] E(\varepsilon_i^2)$$

and

$$(7) \quad E\left[\frac{\sum_{i<j} x_i x_j \varepsilon_i \varepsilon_j}{\sum (x_i^2)^2}\right] = E\left[\frac{\sum_{i<j} x_i x_j}{\sum (x_i^2)^2}\right] \star E(\varepsilon_i\varepsilon_j)$$

assuming $\mathrm{E}(\varepsilon_i^2) = \sigma^2$

$$(8) \quad E\left[\frac{\sum X_i^2}{(\sum X_i^2)^2}\right] E(\varepsilon_i^2) = E\left[\frac{1}{\sum X_i^2}\right]\sigma^2$$

$$\text{ass mi ng } E(\varepsilon_i, \varepsilon_j) = 0$$

$$(8) \quad E\left[\frac{\sum X_i X_j}{(\sum X_i^2)^2}\right] E(\varepsilon_i, \varepsilon_j) = E\left[\frac{\sum X_i X_j}{(\sum X_i^2)^2}\right] * 0$$

then substituting 6,7,8,9 into 5

$$(10) \quad \operatorname{var}(b) = \sigma^2 E\left[\frac{1}{\sum X_i^2}\right]$$

IV.Heteroskedasticity and Autocorrelation

A.Variance-Covariance matrix of disturbances

1.Simple Regression

Data Array $Y = \alpha + \beta X + \varepsilon$ assume $\alpha = 1, \beta = 2$

Obs.No	*Y*	*X*	ε
1	4	1	1
2	3	1	0
3	2	1	–1
4	6	2	1
5	5	2	0
6	4	2	–1
7	8	3	1
8	7	3	0
9	6	3	–1

$$\sum(\varepsilon \mid X = 1) = 0 \quad \frac{\sum[(\varepsilon_1 - \bar{\varepsilon})^2 \mid X = 1]}{n} = \frac{2}{3}$$

$$\sum(\varepsilon \mid X = 2) = 0 \quad \frac{\sum[(\varepsilon_1 - \bar{\varepsilon})^2 \mid X = 2]}{n} = \frac{2}{3}$$

$$\sum(\varepsilon \mid X = 3) = 0 \quad \frac{\sum[(\varepsilon_1 - \bar{\varepsilon})^2 \mid X = 3]}{n} = \frac{2}{3}$$

2.variance-σ*Variance* of disturbances

X	1	2	3
1	$E(\varepsilon_i^2 \mid X=1)$	$E((\varepsilon_i \mid X=2)(\varepsilon_i \mid X=1)$	$E((\varepsilon_i \mid X=3)(\varepsilon_i \mid X=1)$
2	$E((\varepsilon_i \mid X=1)(\varepsilon_i \mid X=2)$	$E(\varepsilon_i^2 \mid X=2)$	$E((\varepsilon_i \mid X=3)(\varepsilon_i \mid X=2)$
3	$E((\varepsilon_i \mid X=1)(\varepsilon_i \mid X=3)$	$E((\varepsilon_i \mid X=2)(\varepsilon_i \mid X=3)$	$E(\varepsilon_i^2 \mid X=3)$

X	1	2	3
	$\frac{2}{3}$	0	0
	0	$\frac{2}{3}$	0
	0	0	$\frac{2}{3}$

The variance-covariance matrix is sometimes written as

X	1	2	3
1	σ_{11}	σ_{21}	σ_{31}
2	σ_{12}	σ_{22}	σ_{32}
3	σ_{13}	σ_{23}	σ_{33}

in the case of the data array above this matrix can be symmetrical where $\sigma^2 = \frac{2}{3}$. These data conform to the classical assumptions since $\mathrm{E}(\varepsilon_i^2) = \sigma^2.\ E(\varepsilon_i, \varepsilon_j) = 0$.

IV. Heteroskedasticity

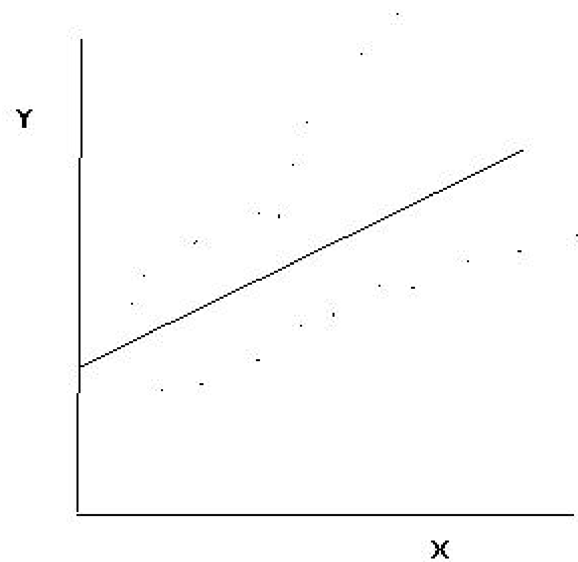
When the disturbance exhibit heteroskedacity, the variance-covariance matrix has unequal variances on the diagonal for example.

s

X	1	2	3
1	σ_{11}	0	0
2	0	σ_{22}	0
3	0	0	σ_{33}

$where \sigma_{11} \neq \sigma_{22} \neq \sigma_{33}$

Diagrammatically, it might look like this


Y
X