Measurement errors in X

We have assumed that all variation is in Y. Measurement error in this variable will not effect the results, as long as they are uncorrelated and unbiased, since they cancel out.

However, we have assumed that X is measured without error, and measurement error in this variable can cause error. Since all error is “vertical", we cannot incorporate this measurement error into our model. As a result, this additional error must, in some way, get incorporated into the model and/or its error.

often it is not true that X is measured without error particularly in meristic relationships

eg.

height of brother $-$ height of sister
body length $-$ scale length
length $-$ weight

Let the measurement error in X_i be denoted as $\delta_i = X_i^* - X_i$

where X_i^* is the measured value and X_i is the true value of the variable.

Then, when fitting the supposed model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

we are actually fitting

$$Y_i = \beta_0 + \beta_1(X_i^* - \delta_i) + \epsilon_i$$

and, multiplying out and grouping variability effects,

$$Y_i = \beta_0 + \beta_1 X_i^* + (\epsilon_i - \delta_i\beta_1)$$

As a result,

a) X_i^* is not fixed (measured without error), it is a random variable

b) The variance term is not longer independent of X_i, since δ_i contains X_i

c) b_0 and b_1 are biased (towards zero) and

lack consistency (ie. $\lim_{n\to\infty} P(\hat{\beta}_i - \beta_i) > \epsilon) = 0$ where ϵ is some arbitrary, positive real number; so $\hat{\beta}_i$ does not tend toward β_i probabilistically as n increases infinitely)

d) There are a couple of cases or aspects of the variation in X_i where variation is not a problem.

a) X_i may be a random variable, not under the control of the investigator. However, this is not a problem as long as the value of X_i is measured without measurement error and is known exactly.

b) the Berkson Model is a special case where measurement error does not effect the results, it cancels out.

In this model, the situation for X_i^* and X_i is reversed. If X_i^* is some fixed value that the investigator is shooting for (for example, by setting some machine value; as a thermostat, adjusting a current speed, or some other machine setting) then the measured value, X_i^* is a constant while the true value, X_i, is the random variable.

What to do?

1) Don't have measurement error.

2) Pretend you don't have measurement error.

3) if only 1 variable has error, then it **must be used as the dependent variable**. Inverse prediction (fitting Y on X and then making inference about X) is the next topic.

4) Measure the measurement error and adjust for it. There are ways to do this.

eg. Snedecor and Cochran (l980): we want to fit

$$Y_i = \beta_0 + \beta_1 X_i,$$

but we are using

$$Y = b_0 + b_1' X' , \quad \text{where } X' \text{ has error so } X' = (X + e)$$

Then b_1' is $\beta_1' = \frac{\beta_1}{1 + \lambda}$

where $\lambda = \frac{\sigma_e^2}{\sigma_x^2}$, indicates the magnitude of the bias

β_1 is regression coefficient for X measured without error and is called the structural regression coefficient. We no longer have an unbiased estimate of this parameter.

obtain estimate of σ_e^2 , error in X = S_e^2

obtain estimate of σ_x^2 , variance in X = S_x^2

assume error in ϵ, e and X is normal then

$$\hat{\lambda} = \frac{S_e^2}{(S_{x'}^2 - S_e^2)}$$

$$\hat{\beta}_1 = b_1 = \frac{b_1'}{(1 + \hat{\lambda})}$$

This estimates with error in Y only (no variability in X) as per the assumptions.

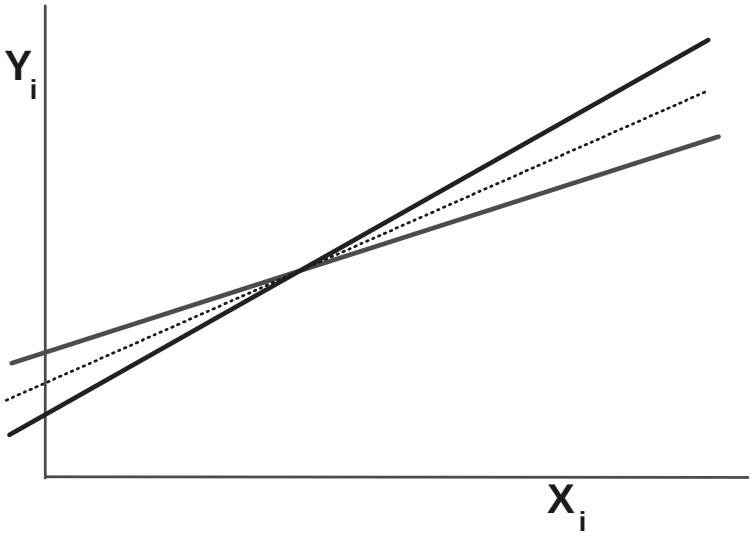
5) another solution – RICKER (1973) approach (limited utility, but individual applications are discussed by RICKER)

Ricker points out that many predictive equations in fisheries are underestimated (bias is actually towards 0, those mentioned have positive slopes)

suggests 2 solutions

(1) Central axis or " geometric mean" axis

Brothers Height (Y)



SISTERS HEIGHT (X)

(a) regress Y on X – vertical error
(b) regress X on Y – horizontal error
(c) for X and Y bivariate normal the line which splits the difference such that

b_1 for Y on X $= \frac{1}{b_1}$ for X on Y

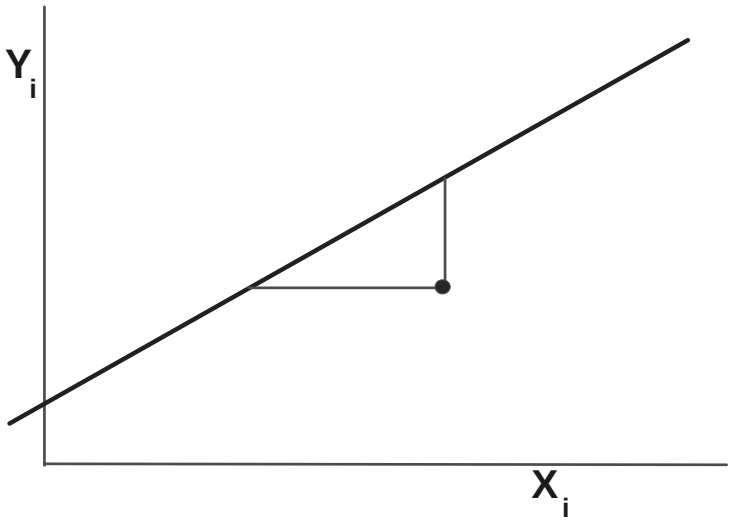
is given by

$$v = \pm \sqrt{\frac{\Sigma y^2}{\Sigma x^2}} = \pm \frac{S_y}{S_x} = \pm \frac{b_1}{r} = \pm \sqrt{b_1^2 + S_{b_1}^2 (N - 2)}$$

(d) Ricker calls the line the GM (geometric mean axis)

THIS LINE IS NOT A BISECTOR, though it will always fall between the other two lines

It is the line is that which minimizes the product of the horizontal and vertical distances of the point from the line.



(e) once the new slope is obtained, the intercept can be calculated as usual

$\bar{Y} - b\prime\bar{X}$

6) MAJOR AXIS – The line that minimizes the SS (perpendicular distance) of observed points to the fitted line

$$z = \frac{\Sigma y^2 - \Sigma x^2 + \sqrt{(\Sigma x^2 - \Sigma y^2)^2 + 4(\Sigma xy)^2}}{2\,\Sigma xy}$$

if e_1 is the measurement error in Y

and e_2 is the measurement error in X

then this equation presumes that $e_1 = e_2$

A MORE GENERAL EQUATION – available in software SUPERCARP
(Wayne Fuller, Stat Dept., Iowa)

employs an expression $\delta = \frac{e_1}{e_2}$, then

$$\hat{\beta}_1 = \frac{\Sigma y^2 - \delta \Sigma x^2 + \sqrt{(\delta \Sigma x^2 - \Sigma y^2)^2 + 4\delta(\Sigma xy)^2}}{2\,\Sigma xy}$$

if $\delta \neq 1$, then the distance minimized is not perpendicular

NOTE: In all of the above cases, once the slope has been calculated, the intercept is obtained by

$$b_0 = \bar{Y} - b_1 \bar{X}$$

For our purposes,

Generally a least squares fit with traditional assumptions will be adequate
We will consider a correction to some equations, particularly

when we expect some theoretical value
eg. $\beta_1 = 3$

Note: Ricker (1973) suggests specific applications of either

– geometric mean axis or

– major axis

and provides equations for the variance of each

Inverse Prediction : Sometimes it is necessary to make predictions of the *independent variable*, X_i, instead of the *dependent variable*, Y_i.

This may occur because we have only the regression equation available, and not the original data.

Or it may be that we are interested in predicting X_i, which is measured without error, from Y_i, which is measured with error.

The process of predicting the independent variable from the dependent variable is called “inverse prediction".

Inverse prediction starts the same as any SLR

The population model is

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

And the usual least squares analysis is done on a sample of n observations from the parent population.

$$\hat{Y}_i = b_0 + {}_1X_i$$

In order to estimate $\hat{X}_i$ for some value of Y_h, we then solve the equation for $\hat{X}_h$.

$$\hat{X}_h = \frac{Y_h - b_0}{b_1} \quad \text{where } b_1 \neq 0 \text{ (ie a relationship must exist)}$$

a confidence interval for the new observation X_h is given by

$$s^2_{X_h} = \frac{MSE}{b_1^2}\left(1 + \frac{1}{n} + \frac{(\hat{X}_h - \bar{X})^2}{\Sigma(X_i - X)^2}\right)$$

which is a transformation of the same equation we used before. This is an approximate value, which your book points out is appropriate if

$$\frac{\left[t_{1-\frac{\alpha}{2},n-2}\right]^2 * MSE}{b_1^2 * \Sigma(X_i - \bar{X})^2} \text{ is small (ie < 0.1)}$$

in our case (Vial example), $\frac{2.306^2 * 2.2}{4^2 * 10} = \frac{11.6988}{160} = 0.07311$, within the recommended

For our vial breakage example;

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	160.00000	160.00000	72.727	0.0001
Error	8	17.60000	2.20000		
C Total	9	177.60000			

Root MSE	1.48324	R-square	0.9009
Dep Mean	14.20000	Adj R-sq	0.8885
C.V.	10.44535		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	10.200000	0.66332496	15.377	0.0001
X	1	4.000000	0.46904158	8.528	0.0001

The equation to predict X is given by, suppose we wish to predict how may transfers would cause 20 vials to be broken.

$$\hat{X}_h = \frac{Y_h - b_0}{b_1} = \frac{Y_h - 10.2}{4} = \frac{20 - 10.2}{4} = \frac{9.8}{4} = 2.45$$

a confidence interval for the new observation $\hat{X}_h$=2.45 is given by

$$s^2_{X_h} = \frac{MSE}{b_1^2}\left(1 + \frac{1}{n} + \frac{(\hat{X}_h - \bar{X})^2}{\Sigma(X_i - X)^2}\right)$$

$$= \frac{2.2}{4^2}\left(1 + \frac{1}{10} + \frac{(2.45-1)^2}{20 - \frac{10^2}{10}}\right) = 0.1375\left(\frac{1}{10} + \frac{2.1025}{10}\right)$$

$$= 0.1375 * 0.31025 = 0.04266$$

$$s_{X_h} = \sqrt{0.04266} = 0.20654$$

since $t_{\frac{\alpha}{2},\,8\,df} = 2.306$, then

$$P(\hat{X}_{Y=20} - t_{1-\frac{\alpha}{2},n-2}\, s_{\hat{X}_h} \leq E(\hat{X}) \leq \hat{X}_{Y=20} + t_{1-\frac{\alpha}{2},n-2}\, s_{\hat{X}_h}) = 1\text{-}\alpha$$

$$P(2.45 - 2.306*0.2065 \leq E(\hat{X}) \leq 2.45 + 2.306*0.2065) = 0.95$$

$$P(1.9738 \leq E(\hat{X}) \leq 2.9262) = 0.95$$

so it appears most likely that 2 transfers would be involved in this damage, though 3 transfers is not out of the question.