Measurement Error

*Definitions***.** For two variables, X and Y, the following hold:

Parameter	**Explanation**
$E(X) = \frac{\sum X_i}{N} = \mu_x$	Expectation, or Mean, of X
$V(X) = E[(X - \mu_x)^2] = \sigma_x^2$	Variance of X
$SD(X) = \sqrt{V(X)} = \sigma_x$	Standard Deviation of X
$COV(X,Y) = E[(X - \mu_x)(Y - \mu_y)] = \sigma_{xy}$	Covariance of X and Y
$CORR(X,Y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = r_{xy} = r_{yx}$	Correlation of X and Y
$\beta_{yx} = \frac{\sigma_{xy}}{\sigma_x^2}$	Slope coefficient for the Bivariate regression of Y on X (Y dependent)

Question: Suppose X suffers from random measurement error - that is, the values of X that we observe differ randomly from the true values that we are interested in. For example, we might be interested in income. Since people do not remember their income exactly, reported income will sometimes be higher and sometimes be lower than true income. In such a case, how does random measurement error affect the various statistical measures we are typically interested in? That is, how does unreliability affect our statistical measures and conclusions?

Revised Question: Let us put the question more formally. Let $X = X_t + \varepsilon$, where ε is a random error term (i.e. has mean 0 and variance s^2_ε). That is, X_t is the "true" value of the variable, and X is the flawed measure of the variable that is observed. We want to see how the statistics for the observed variable, X, differ from the statistics for the true variable, X_t. When thinking about this question, keep in mind that, because ε is a random error term, it is independent from all other variables (except itself), e.g. $COV(X, \varepsilon) = COV(Y, \varepsilon) = 0$.

Definition of Reliability: The reliability of a variable is defined as:

$$REL(X) = \frac{\sigma_{x_t}^2}{\sigma_x^2} = r^2_{XtX}$$

The first equality says reliability is true variance divided by total variance. The second equality says the reliability of a variable is the squared correlation between the true value of the variable and the observed value that suffers from random measurement error. If there is no random measurement error, reliability = 1.

*Some additional rules for expectations***.** Before answering the question, the following additional rules are helpful. Let A, B, C, and D be random variables. Then,

(1) E(A + B) = E(A) + E(B)
(2) If A and B are independent, V(A + B) = V(A) + V(B)
(3) COV(A + B, C + D) = COV(A,C) + COV(A,D) + COV(B,C) + COV(B,D)

Hypothetical Data. To help illustrate the points that will follow, we create a data set where the true measures (Yt and Xt) have a correlation of .7 with each other – but the observed measures (Y and X) both have some degree of random measurement error, and the reliability of both is .64. The way I am constructing the data set, using the `corr2data` command, there will be no sampling variability, i.e. we can act as though we have the entire population.

```
. matrix input corr = (1,.7,0,0\.7 ,1,0,0\0,0,1,0\0,0,0,1)
. matrix input sd = (4,8,3,6)
. matrix input mean = (10,7,0,0)
. corr2data Yt Xt ey ex, corr(corr) sd(sd) mean(mean) n(500)
 (obs 500)
. * Create flawed measures with random measurement error
. gen Y = Yt + ey
. gen X = Xt + ex
```

Effects of Unreliability

A. For the mean:

$E(X) = E(X_t + \varepsilon) = E(X_t) + E(\varepsilon) = E(X_t)$ [Expectations rule 1]

NOTE: Remember, since errors are random, ε has mean 0.

Implication: Random measurement error does not bias the expected value of a variable - that is, $E(X) = E(X_t)$

B. For the variance:

$V(X) = V(X_t + \varepsilon) = V(X_t) + V(\varepsilon)$ [Expectations rule 2]

NOTE: Remember, $COV(X_t, \varepsilon) = 0$ because ε is a random disturbance.

Implication: Random measurement error does result in biased variances. The variance of the observed variable will be greater than the true variance.

A & B illustrated with our hypothetical data. We see that the flawed, observed measures have the same means as the true measures – but their variances & standard deviations are larger:

```
. sum  Yt Y Xt X

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          Yt |       500          10           4  -2.639851   22.83863
           Y |       500          10           5  -3.706503   26.55569
          Xt |       500           7           8  -16.16331   28.80884
           X |       500           7          10  -23.81675   38.49127
```

C. For the covariance (we'll let Yt stand for the perfectly measured Y variable):

$COV(X,Yt) = COV(X_t + \varepsilon, Yt) = COV(X_t, Yt) + COV(\varepsilon, Yt)$

$= COV(X_t, Yt)$ [Expectations rule 3]

NOTE: Remember, $COV(\varepsilon,Yt) = 0$ because ε is a random disturbance.

Implication: Covariances are <u>not biased</u> by random measurement error.

C illustrated with our hypothetical data. Random measurement error in X does NOT affect the covariance:

```
. corr Yt Xt X, cov

(obs=500)

             |       Yt       Xt        X
-------------+---------------------------
          Yt |       16
          Xt |     22.4       64
           X |     22.4       64      100
```

D. For the correlation:

$$r_{xy_t} = \frac{\sigma_{XY_t}}{\sigma_X \sigma_{Y_t}} \text{ , } r_{x_t y_t} = \frac{\sigma_{X_t Y_t}}{\sigma_{X_t} \sigma_{Y_t}} = \frac{\sigma_{XY_t}}{\sigma_{X_t} \sigma_{Y_t}} \text{ .}$$

Thus, when X and Yt covary positively, $CORR(X,Yt) \le CORR(X_t,Yt)$

Implication: Random measurement error produces a <u>downward bias</u> in the bivariate correlation. This is often referred to as <u>attenuation</u>.

D with hypothetical data. The correlation is attenuated by random measurement error:

```
. corr Yt Xt X

(obs=500)

             |       Yt       Xt        X
-------------+---------------------------
          Yt |   1.0000
          Xt |   0.7000   1.0000
           X |   0.5600   0.8000   1.0000
```

Note that the correlation between X and Xt is .8 – and that the correlation between X and Yt (.56) is only .8 times as large as the correlation between Xt and Yt (.7). Also, the .8 correlation between X and Xt means that the reliability of X is .64.

E. For $ß_{YtX}$: (Yt is perfectly measured, X has random measurement error)

$$\beta_{YtX} = \frac{\sigma_{XYt}}{\sigma^2_X},\ \beta_{YtXt} = \frac{\sigma_{XtYt}}{\sigma^2_{Xt}}$$ Thus, when X and Yt covary positively, $ß_{YtX} \leq ß_{YtXt}$

Implication: Random measurement error in the Independent variable produces a downward bias in the bivariate regression slope coefficient.

E with hypothetical data. In a bivariate regression, random measurement error in X causes the slope coefficient to be attenuated, i.e. smaller in magnitude. First we run the regression between the true measures, and then we run the regression of Yt with the flawed measure X:

```
. reg Yt Xt

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  1,   498) =  478.47
       Model |  3912.16007     1  3912.16007           Prob > F      =  0.0000
    Residual |  4071.84001   498  8.17638555           R-squared     =  0.4900
-------------+------------------------------           Adj R-squared =  0.4890
       Total |  7984.00008   499  16.0000002           Root MSE      =  2.8594

------------------------------------------------------------------------------
          Yt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          Xt |        .35   .0160008    21.87   0.000     .3185627    .3814373
       _cons |       7.55    .169994    44.41   0.000     7.216006    7.883994
------------------------------------------------------------------------------

. reg Yt X

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  1,   498) =  227.52
       Model |  2503.78247     1  2503.78247           Prob > F      =  0.0000
    Residual |  5480.21761   498   11.004453           R-squared     =  0.3136
-------------+------------------------------           Adj R-squared =  0.3122
       Total |  7984.00008   499  16.0000002           Root MSE      =  3.3173

------------------------------------------------------------------------------
          Yt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           X |       .224   .0148503    15.08   0.000     .1948231    .2531769
       _cons |      8.432   .1811488    46.55   0.000      8.07609     8.78791
------------------------------------------------------------------------------
```

Note that X has a reliability of .64 – and the slope coefficient using the flawed X (.224) is only .64 times as large as the slope coefficient using the perfectly measured Xt (.35).

F. For $ß_{YXt}$: (Now Y is measured with random error, while Xt is measured perfectly)

$$\beta_{YXt} = \frac{\sigma_{X_tY}}{\sigma^2_{X_t}}, \quad \beta_{Y_tX_t} = \frac{\sigma_{X_tY_t}}{\sigma^2_{X_t}}$$. Thus, $ß_{YXt} = ß_{YtXt}$

Implication: Random measurement error in the Dependent variable does not bias the slope coefficient. HOWEVER, it does lead to larger standard errors. Recall that the formula for the standard error of b is

$$s_b = \sqrt{\frac{1-R^2}{(N-K-1)}} * \frac{s_Y}{s_X}$$

When you have random measurement error in Y, R^2 goes down because of the previously noted downward bias. This increases the numerator. Also, the variance of Y goes up, which further increases the standard error.

F with hypothetical data. Random measurement error in Y does not cause the slope coefficient to be biased – but it does cause the standard error for the slope coefficient to be larger and the t value smaller. Again we run the true regression followed by the regression of Y with Xt.

```
. reg Yt Xt

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  1,   498) =  478.47
       Model |  3912.16007     1  3912.16007           Prob > F      =  0.0000
    Residual |  4071.84001   498  8.17638555           R-squared     =  0.4900
-------------+------------------------------           Adj R-squared =  0.4890
       Total |  7984.00008   499  16.0000002           Root MSE      =  2.8594

------------------------------------------------------------------------------
          Yt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          Xt |        .35   .0160008    21.87   0.000     .3185627    .3814373
       _cons |       7.55    .169994    44.41   0.000     7.216006    7.883994
------------------------------------------------------------------------------

. reg Y Xt

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  1,   498) =  227.52
       Model |  3912.16001     1  3912.16001           Prob > F      =  0.0000
    Residual |  8562.84011   498   17.194458           R-squared     =  0.3136
-------------+------------------------------           Adj R-squared =  0.3122
       Total |  12475.0001   499  25.0000002           Root MSE      =  4.1466

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          Xt |        .35   .0232035    15.08   0.000     .3044111    .3955889
       _cons |       7.55   .2465171    30.63   0.000     7.065658    8.034342
```

Additional implications

- When you have more than one independent variable, random measurement error can cause coefficients to be biased either upward or downward. As you add more variables to the model, all you can really be sure of is that, if the variables suffer from random measurement error (and most do) the results will probably be at least a little wrong!
- Reliability is a function of both the total variance and the error variance. True variance is a population characteristic; error variance is a characteristic of the measuring instrument. As Wiley and Wiley note, an item rarely has a "unique" reliability. Any social process that tends to increase consensus results in a reduction in the true variability in attitudes without necessarily affecting the characteristics of the measuring instrument. An increase in reliability can always be achieved by conducting measurements on a more heterogeneous population.
- Comparisons of any sort can be distorted by differential reliability of variables. For example, if comparing effects of two variables, one variable may appear to have a stronger effect simply because it is better measured. If comparing, say, husbands and wives, the spouse who gives more accurate information may appear more influential. For a more detailed discussion of how measurement error can affect group comparisons, see Thomson, Elizabeth and Richard Williams (1982) "Beyond wives family sociology: a method for analyzing couple data" Journal of Marriage and the Family Vol 44 999:1008
- The fact that reliabilities differ between groups does not necessarily mean that one group is more "accurate." It may just mean that there is less true variance in one group than there is in another. Examining error variances is therefore often more informative than just looking at reliabilities.

Dealing with measurement error. For the most part, this is a subject for a research methods class or a more advanced statistics class. I'll toss out a few ideas for now:

- Collect better quality data in the first place. Make questions as clear as possible.
- Measure multiple indicators of concepts. When more than one question measures a concept, it is possible to estimate reliability and to take corrective action. For a more detailed discussion on measuring reliability, see Reliability and Validity Assessment, by Edward G. Carmines and Richard A. Zeller. 1979. Paper # 17 in the Sage Series on Quantitative Applications in the Social Sciences. Beverly Hills, CA: Sage.
- Create scales from multiple indicators of a concept. The scales will generally be more reliable than any single item would be. In SPSS you might use the FACTOR or RELIABILITY commands; in Stata relevant commands include `factor` and `alpha`.
- Use advanced techniques, such as LISREL, which let you incorporate multiple indicators of a concept in your model. Ideally, LISREL "purges" the items of measurement error hence producing unbiased estimates of structural parameters. In Stata 12, this can also be done with the new `sem` (Structural Equation Modeling) command.