

Measurement: *Units and Errors*

- ***Units of Measurement: Introduction***
- ***SLR Application***
- ***Example: Bodyfat***
- ***Beta Regressions (w/ Standardized variables)***
- ***Assessing Meaningfulness: Beta Regressions v. Elasticities***
- ***Measurement Errors (Errors in Variables): Introduction***
- ***SLR Application: The Classic Mismeasurement Assumptions***
- ***... Errors on the LHS***
- ***... on the RHS***
- ***... and on Both Sides***

Units of Measurement: Introduction

- 1) Units, units units... tens, hundreds, thousands, millions, gazillions... inches, feet, miles, millimeters, centimeters, meters, kilometers, light years... ounces, pounds, tons, grams, kilograms, megatonnes, gigatonnes... degrees Fahrenheit, Centigrade, Celsius, Kelvin... etc etc etc.
- 2) Units of measurement are often *ad hoc* and a matter of convenience. And so you might wonder: Does econometrics care? Do regression results vary with units of measurement? Inquiring minds want to know!
- 3) Not to be too flip, but the short answer is: regression results that are sensitive to units of measurement will vary with units of measurement... and those that are not, will not.
- 4) But importantly, the results that really matter will not be impacted by changes in the units of measurement. And so perhaps not surprisingly, changes in units of measurement will have no material impact on the results from OLS estimation of SLR or MLR models. So feel free to rescale to your hearts content.

Let's illustrate with some SLR models.

| Customary Units of Measurement - Chart | | | |
|--|-----------------|-------------------|-----------------|
| Length | Weight | Capacity | Time |
| 12 in = 1 ft | 16 oz = 1 lb | 128 fl oz = 1 gal | 60 sec = 1 min |
| 3 ft = 1 yrd | 2000 lb = 1 ton | 2 pt = 1 qt | 60 min = 1 hr |
| 5,280 ft = 1 mi | | 8 pt = 1 gl | 24 hr = 1 day |
| 1,760 yrd = 1 mi | | 4 qt = 1 gal | 7 days = 1 wk |
| | | | 52 wk = 1 yr |
| | | | 12 mon = 1 yr |
| | | | 365 days = 1 yr |

Measurement: Units and Errors

SLR Application

5) Consider a traditional SLR model, SLR 1, in which y is regressed on x . From before, you know that included in the regression results are:

- a) SRF₁: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, with the slope coefficient defined by $\hat{\beta}_1 = \rho_{xy} \frac{S_y}{S_x}$,
- b) $R_1^2 = \rho_{xy}^2$ (the subscript on R_1^2 reflects the Model #, SLR 1), and
- c) $t_x^2 = (n-2) \frac{(1-R_1^2)}{R_1^2}$ (the square of the t stat for the slope coefficient).

6) Now consider linear rescalings of x and y , allowing for possibly different intercepts and slopes: $w = a + bx$ and $z = c + dy$. Since rescalings should preserve the ordering of outcomes, b and d are both positive).

- a) Note that rescalings can include changes in units of measurement (the slope coefficients in the rescalings, b and d above), as well as shifts in origins (the constant term in the rescalings, a and c above).
- b) Here's an example: Degrees Fahrenheit are just a rescaling of degrees Centigrade
 - i) $tempF = 32 + 2.12 \cdot tempC$
(temperature in Fahrenheit is 2.12 times temperature in Celsius + 32 degrees).
 - ii) In this rescaling example there is a shift in origin (+32) as well as a change in the units of measurement (2.12).

7) Recall that:

- a) Sample correlations are unaffected by rescaling: $\rho_{wz} = \frac{S_{wz}}{S_w S_z} = \rho_{xy} = \frac{S_{xy}}{S_x S_y}$, and
- b) standard deviations will reflect the change in units: $S_w = b S_x$ and $S_z = d S_y$.

8) Now estimate Model SLR 2, working with the rescaled data and regressing z on w . Among the results are:

- a) SRF₂: the new slope coefficient: $\rho_{wz} \frac{S_z}{S_w} = \rho_{xy} \frac{d S_y}{b S_x} = \hat{\beta}_1 \frac{d}{b} \dots$ (**impact!**),
- b) the new: $R_2^2 = \rho_{wz}^2 = \rho_{xy}^2 = R_1^2 \dots$ (**no impact!**), and
- c) the new t stat: $t_w^2 = (n-2) \frac{(1-R_2^2)}{R_2^2} = (n-2) \frac{(1-R_1^2)}{R_1^2} = t_x^2 \dots$ (**no impact!**).

9) And so:

- a) **Slope coefficient:** The estimated slope coefficient will be impacted by the rescaling factors b and d . If the y values are multiplied by a factor of 10 (so $d = 10$), the estimated slope coefficients will increase by the same factor of 10. And if instead the values of the

Measurement: *Units and Errors*

RHS variable are multiplied by a factor of 10 (so $b = 10$), then new slope coefficient will be reduced by 90%... with a magnitude that is 10% (one tenth) of the magnitude of the previously estimated slope coefficient.

- b) **R-sq**: R squared is unaffected by the rescaling, since the correlation of the rescaled variables is the same as the correlation of the original variables.
- c) **t stat**: t statistics are also unaffected by the rescaling since the number of observations is unchanged and **R-sq** is unchanged. Since the *dofs* are unchanged, the associated p values are unchanged by the rescaling, as is statistical significance. Notice that since the t stats are unchanged and since $t_x = \frac{\hat{\beta}_x}{std\ err_x}$, the change in standard errors basically unwinds whatever impact the rescaling had on the estimated coefficients. So standard errors will be impacted by the rescaling of the variables in the same way that estimated slope coefficients were.

- 10) So in general, rescaling will impact the estimated coefficients, SSRs, SSTs, SSEs, MSE/RMSEs and standard errors. But it will have no real impact on the estimated model as t stats, p values, and statistical significance as well as R-sq's and adj. R-sq's are unaffected by the rescaling of variables.
- 11) **Important takeaway**: Students (but not you, of course!) often believe that the magnitude of an estimated OLS coefficient (SLR or MLR) tells them something. Perhaps... but the truth is that you can make that magnitude anything you want it to be by just rescaling the variables in the model. So pay attention to the signs of coefficients and things like t stats, p value statistical significance, R-sq's and adj. R-sq's ... but don't be deceived by coefficient estimates that are thought to be large or small. Estimated coefficient magnitudes are completely dependent on units of measurement.

Example: *Bodyfat*

12) Here's an example using the bodyfat dataset.

- a) In that dataset, there are two measure of bodyfat, *Brozek* and *Siri*, which are in turn, functions of *density*. These two measures are related by a simple rescaling:
 - i) $Brozek = (457 / density) - 414.2$, and
 - ii) $Siri = (495 / density) - 450$, and so
 - iii) $Siri = (495 / 457) Brozek - (450 - (495 / 457) 414.2) = 1.08 \cdot Brozek - 1.36$
- b) As well, height can be measured in inches (*hgt*) or in meters (*hgt_m*)... and weight can be measured in pounds (*wgt*) or kilos (*wgt_kg*).

Measurement: *Units and Errors*

c) Here are regression results involving rescalings of all variables:

```
. reg Brozek hgt wgt
```

| Source | SS | df | MS | Number of obs | = | 252 |
|----------|------------|-----|------------|---------------|---|--------|
| Model | 6829.45017 | 2 | 3414.72509 | F(2, 249) | = | 99.92 |
| Residual | 8509.70341 | 249 | 34.1755157 | Prob > F | = | 0.0000 |
| | | | | R-squared | = | 0.4452 |
| | | | | Adj R-squared | = | 0.4408 |
| Total | 15339.1536 | 251 | 61.1121657 | Root MSE | = | 5.846 |

| Brozek | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] | |
|--------|-----------|-----------|-------|-------|----------------------|-----------|
| hgt | -.6722339 | .1058974 | -6.35 | 0.000 | -.8808027 | -.4636652 |
| wgt | .1844145 | .0131983 | 13.97 | 0.000 | .15842 | .210409 |
| _cons | 33.04038 | 7.076723 | 4.67 | 0.000 | 19.10252 | 46.97825 |

```
. reg Siri hgt_m wgt_kg
```

| Source | SS | df | MS | Number of obs | = | 252 |
|----------|------------|-----|------------|---------------|---|--------|
| Model | 8012.4211 | 2 | 4006.21055 | F(2, 249) | = | 99.92 |
| Residual | 9983.72057 | 249 | 40.0952633 | Prob > F | = | 0.0000 |
| | | | | R-squared | = | 0.4452 |
| | | | | Adj R-squared | = | 0.4408 |
| Total | 17996.1417 | 251 | 71.6977756 | Root MSE | = | 6.3321 |

| Siri | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] | |
|--------|-----------|-----------|-------|-------|----------------------|-----------|
| hgt_m | -28.66657 | 4.515859 | -6.35 | 0.000 | -37.56072 | -19.77242 |
| wgt_kg | .440371 | .0315167 | 13.97 | 0.000 | .3782976 | .5024444 |
| _cons | 34.42887 | 7.665159 | 4.49 | 0.000 | 19.33205 | 49.52568 |

As advertised, while there are a number of differences in the regression results, t stats, p values, statistical significance, R^2 's and \bar{R}^2 's are not impacted by the rescaling of the different variables.¹

Beta Regressions (w/ standardized variables)

13) *Beta Regressions* involve standardized variables and accordingly, provide us with a good example of rescaling in action. To run a beta regression, all of the variables are first standardized to have mean zero and variance one, and then OLS estimation is employed to estimate the unknown parameter values. As you'll see below, such regressions are especially useful in assessing economic significance, or *meaningfulness*, and provide an attractive alternative to elasticities in that regard.

¹ Note that t stats and p values for the estimated constant coefficient are not invariant to rescaling, so long as there is a change of origin in the rescaling.

Measurement: *Units and Errors*

14) Consider the standard SLR model with dependent variable y and RHS variable x .

- a) To run the beta regression of y on x , you first standardize the variables to have mean 0 and variance 1:

$$y_i^* = \frac{y_i - \bar{y}}{S_y} \text{ and } x_i^* = \frac{x_i - \bar{x}}{S_x}.$$

Note that standardization is a linear rescaling of the variables (for the y 's, the rescaling has a constant term of $-\frac{\bar{y}}{S_y}$ and a slope of $\frac{1}{S_y}$).

- b) To run the beta regression of y on x , just regress the standardized variables on one another: regress y^* on x^* .
- c) Looking at the results from the beta regression, you'll see that the estimated slope coefficients are just the correlation of the LHS and RHS variables (with or without standardization, it doesn't matter) and the estimated intercept is 0:
- i) Since the standard deviations of the standardized variables are both 1, the estimated slope coefficient is the correlation between x^* and y^* , $\rho_{x^*y^*}$. This will be the same as the correlation between x and y , ρ_{xy} , since x^* and y^* are just linear rescalings of x and y , and correlations are unaffected by such rescalings.
- ii) And since $\bar{y}^* = 0$ and $\bar{x}^* = 0$ by construction, the estimated intercept coefficient is 0 since $\bar{y}^* - \rho_{x^*y^*}\bar{x}^* = 0$.

15) Beta regression are insensitive to linear rescalings, because the standardization process unwinds the effects of rescaling. Don't believe me? Consider the y 's above.

a) Standardized y_i : $y_i^* = \frac{y_i - \bar{y}}{S_y}$

- b) Rescale and standardize: Define the z 's to be rescaled y 's: $z = c + dy$. Since $\bar{z} = c + d\bar{y}$ and $S_z = dS_y$ (since $d > 0$), the standardized z_i is:

$$\frac{z_i - \bar{z}}{S_z} = \frac{(c + dy_i) - (c + d\bar{y})}{dS_y} = \frac{d(y_i - \bar{y})}{dS_y} = \frac{(y_i - \bar{y})}{S_y}$$

... the same as the standardized y_i !

Measurement: *Units and Errors*

- 16) Example cont'd: Continuing the bodyfat example started above. To run the beta regression you just add ", *beta*" (no quotes) to the end of the regression command.

```
. reg Brozek hgt wgt, beta
```

| Source | SS | df | MS | Number of obs | = | 252 |
|----------|------------|-----|------------|---------------|---|--------|
| Model | 6829.45017 | 2 | 3414.72509 | F(2, 249) | = | 99.92 |
| Residual | 8509.70341 | 249 | 34.1755157 | Prob > F | = | 0.0000 |
| | | | | R-squared | = | 0.4452 |
| | | | | Adj R-squared | = | 0.4408 |
| Total | 15339.1536 | 251 | 61.1121657 | Root MSE | = | 5.846 |

| Brozek | Coef. | Std. Err. | t | P> t | Beta |
|--------|-----------|-----------|-------|-------|------------------|
| hgt | -.6722339 | .1058974 | -6.35 | 0.000 | <u>-.3149752</u> |
| wgt | .1844145 | .0131983 | 13.97 | 0.000 | <u>.6932955</u> |
| _cons | 33.04038 | 7.076723 | 4.67 | 0.000 | . |

- 17) Note that the OLS/MLR coefficients, Std. Err.'s, t's and p values are displayed in the usual places... the beta coefficients are on the right side of the output. As anticipated, the *cons* coefficient is zero in the beta regression.
- 18) The following shows that beta regressions are in fact regressions with standardized variables.

```
. *Standardize the Variables
.
. egen muBrozek = mean(Brozek)
. egen sdBrozek = sd(Brozek)
. gen zBrozek = (Brozek-muBrozek)/sdBrozek
.
. egen muhgt = mean(hgt)
. egen sdhgt = sd(hgt)
. gen zhgt = (hgt-muhgt)/sdhgt
.
. egen muwgt = mean(wgt)
. egen sdwgt = sd(wgt)
. gen zwgt = (wgt-muwgt)/sdwgt
.
. reg zBrozek zhgt zwgt
```

| Source | SS | df | MS | Number of obs | = | 252 |
|----------|------------|-----|------------|---------------|---|--------|
| Model | 111.752707 | 2 | 55.8763533 | F(2, 249) | = | 99.92 |
| Residual | 139.247286 | 249 | .559226047 | Prob > F | = | 0.0000 |
| | | | | R-squared | = | 0.4452 |
| | | | | Adj R-squared | = | 0.4408 |
| Total | 250.999992 | 251 | .99999997 | Root MSE | = | .74781 |

| zBrozek | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] |
|---------|------------------|-----------|-------|-------|----------------------|
| zhgt | <u>-.3149752</u> | .0496182 | -6.35 | 0.000 | -.4127001 -.2172503 |
| zwgt | <u>.6932955</u> | .0496182 | 13.97 | 0.000 | .5955706 .7910204 |
| _cons | <u>-1.25e-07</u> | .0471079 | -0.00 | 1.000 | -.0927808 .0927806 |

As anticipated, the coefficients in the standardized variables are exactly the same as the beta coefficients in the beta regression above.

Measurement: *Units and Errors*

19) Finally, here are the results from the Siri regression that we considered earlier. Note that as anticipated, the beta regression coefficients in this model are exactly the same as the beta regression coefficients in the Brozek model, in spite of the fact that all of the variables have been rescaled in the Siri model.

20) And so as promised, beta regression coefficients are insensitive to linear rescalings of the data.

```
.  
. reg Siri hgt_m wgt_kg, beta
```

| Source | SS | df | MS | Number of obs | = | 252 |
|----------|------------|-----|------------|---------------|---|--------|
| Model | 8012.4211 | 2 | 4006.21055 | F(2, 249) | = | 99.92 |
| Residual | 9983.72057 | 249 | 40.0952633 | Prob > F | = | 0.0000 |
| Total | 17996.1417 | 251 | 71.6977756 | R-squared | = | 0.4452 |
| | | | | Adj R-squared | = | 0.4408 |
| | | | | Root MSE | = | 6.3321 |

| Siri | Coef. | Std. Err. | t | P> t | Beta |
|--------|-----------|-----------|-------|-------|-----------|
| hgt_m | -28.66657 | 4.515859 | -6.35 | 0.000 | -.3149753 |
| wgt_kg | .440371 | .0315167 | 13.97 | 0.000 | .6932955 |
| _cons | 34.42887 | 7.665159 | 4.49 | 0.000 | . |

Assessing Meaningfulness: Beta Regressions v. Elasticities

21) We have previously used elasticities to assess economic significance, on the argument that elasticities are insensitive to changes in the units of scale/measurement. That is true; they are insensitive to units of measurement. But elasticities are not insensitive to all rescalings, as they will typically be affected by changes in origin.

22) Here's an example, working with data from the NY Auto Club (taken from an exercise from the Biostatistics and Epidemiology Department at UMass Amherst).²

² Sources:

<http://people.umass.edu/biep640w/pdf/simple%20linear%20regression%20ny%20auto%20club%20STATA.pdf>
and <http://people.umass.edu/biep640w/webpages/regression.htm>.

Measurement: *Units and Errors*

Example: *Emergency Calls to the New York Auto Club*

- a) The data consist of 28 daily observations recording the daily high temperature measured in degrees Fahrenheit, *highf*, and the number of emergency calls to the New York Auto Club, *calls*, during the last two weeks in January of 1993 and 1994.
- b) I generate *highc*, the high temperature in degrees Celsius, estimate several simple SLR models, and report the associated elasticities:

Fahrenheit:

```
. reg calls highf
. margins, eyex(_all) atmean
```

| | | Delta-method | | | | |
|--|-------|------------------|-----------|-------|-------|------------------------|
| | | ey/ex | Std. Err. | t | P> t | [95% Conf. Interval] |
| | highf | <u>-1.043587</u> | .3611748 | -2.89 | 0.008 | -1.785992 - .3011816 |

.
Celsius:

```
. reg calls highc
. margins, eyex(_all) atmean
```

| | | Delta-method | | | | |
|--|-------|------------------|-----------|-------|-------|-----------------------|
| | | ey/ex | Std. Err. | t | P> t | [95% Conf. Interval] |
| | highc | <u>-.1522105</u> | .0526785 | -2.89 | 0.008 | -.2604927 -.0439283 |

.
Regression results:

| | (1) calls | (2) calls |
|-----------|---------------------|----------------------|
| high | -120.3** (-3.03) | |
| highc | | -255.0** (-3.03) |
| _cons | 8825.7*** (5.68) | 4976.1*** (10.03) |
| N | 28 | 28 |
| R-sq | 0.261 | 0.261 |
| adj. R-sq | 0.232 | 0.232 |

t statistics in parentheses

- c) As expected the t stats, p values, statistical significance, R^2 's and \bar{R}^2 's are not impacted by the rescaling of temperature. But while the elasticities are both negative, they have very different magnitudes (the Fahrenheit effect is 8x the Celsius effect), and are clearly not insensitive to the rescaling.

Elasticities: *highf* (Fahrenheit): -1.04 and *highc* (Celsius): -.15

Measurement: *Units and Errors*

- d) So as advertised, elasticities are not invariant to rescalings involving changes in origin.
- 23) In contrast, beta regression coefficients are insensitive to all rescalings, including changes in units of measurement as well as changes in origin. This invariance makes them an attractive alternative measure of economic significance, or meaningfulness.
- 24) So we have two measures of economic significance: elasticities and beta regressions. These are the two most common measures of meaningfulness with which I'm familiar.
- a) **Elasticities**: connect percentage changes in the x 's with percentage changes in the predicted y 's; if the elasticity is, say 0.5, then (starting at the means) a 10% increase in x is associated with a 5% increase in predicted y .
- b) **Beta regressions**: connect changes in the x 's (measured in standard deviations) and with changes in the predicted y 's (also measured in standard deviations); if the beta regression coefficient is again, say 0.5, then a 10% standard deviation increase in x is associated with a 5% standard deviation increase in predicted y .
- 25) **Any Agreement?** You might ask: Are these two measures of significance often in agreement? That question can arise in at least two ways:
- a) *Way 1*: In the context of looking at a single coefficient and asking: Is that estimate meaningful, or economically significant?
- b) *Way 2*: In the context of MLR models in which you are looking across the estimated coefficients and trying to understand which estimated incremental effect is more meaningful, or economically significant.
- 26) I think it's fair to say that you typically get similar results with the two approaches... but as you'll see below, that is certainly not guaranteed.
- a) The following tables present elasticities and beta coefficients for two models with which you are familiar. The first table shows results from a typical sovereign debt model; the second shows results for a typical bodyfat model. You'll see that in the sovereign debt analysis, the beta coefficients and elasticities tell roughly similar stories. But that is most definitely not the case in the bodyfat example. In that example, *hgt* has greater economic significance using the elasticity metric, while *wgt* wins the day with beta regression.

| | Elasticities | Beta Coeffs. |
|-------------|---------------------|---------------------|
| corrupt | 0.442 | 0.702 |
| lngdp | 0.262 | 0.371 |
| | | |
| debt_gdp | -0.073 | -0.168 |
| inflation | -0.035 | -0.118 |
| deficit_gdp | -0.026 | -0.140 |

| | Elasticities | Beta Coeffs. |
|-----|---------------------|---------------------|
| hgt | -2.498 | -0.315 |
| wgt | 1.748 | 0.693 |

- b) So you can't assume that the two approaches are always consistent. Good practice will have you looking at both metrics, and seeing what each tells you.

Measurement: *Units and Errors*

Measurement Error (Errors in Variables): Introduction

- 27) More or less sizable, mismeasurement (sometimes called *errors in variables*) is pervasive. But does it matter? Will variable measurement errors impact OLS coefficient estimates, standard errors, t stats, statistical significance, etc etc. Inquiring minds want to know!
- 28) Not to be too flip, but the short answer is: Measurement error matters when it matters, and otherwise doesn't matter. In a few specific cases (discussed below), we can say something about the impact of measurement error. But in general, we need to know about the specifics of the measurement error to say anything about the OLS impact..
- 29) Suppose that we can say something about, say, whether the magnitudes of OLS estimates increase or decrease, or perhaps are unchanged given measurement errors in variables. As with omitted variable impact/bias, sometimes just knowing the direction/nature of the impact can be useful:
- a) So that you might say something like: *The estimated effect is large in magnitude, and would be even larger if not for the measurement errors in variables.*
 - b) Of course, you might also be forced to say something like: *I know not to trust my OLS estimates, because I've got lots of measurement error and I have no idea how those errors are impacting coefficient estimates, t stats and so forth.*

Let's illustrate with some SLR models.

SLR Application: The Classic Mismeasurement Assumptions

- 30) Consider a traditional OLS/SLR model, in which you have data on the independent RHS variable, x , and the dependent LHS variable y , and are using OLS to estimate a linear relationship between the two variables: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.
- 31) Now suppose that the data you are working with is subject to measurement error, so that:
- $$y_i = y_i^* + \Delta y_i \text{ and } x_i = x_i^* + \Delta x_i,$$
- where the y_i^* 's and x_i^* 's are the true values of the two variables and the Δy_i 's and Δx_i 's are the measurement errors added in when the data are generated.
- 32) You'd like to be working with the true data, the y_i^* 's and x_i^* 's, and estimating the parameters in the linear relationship between those two variables. But that will never happen, because you don't have that data.
- 33) Recall that for OLS/SLR models, and when regressing y on x , the estimated slope coefficient, $\hat{\beta}_1$, will be defined by $\hat{\beta}_1 = \frac{S_{yx}}{S_{xx}}$. And if we had the true data, we would regress y^* on x^* ,
- $$\text{and the estimated slope would be } \hat{\beta}_1^* = \frac{S_{y^*x^*}}{S_{x^*x^*}}.$$
- 34) The question is thus: How will the estimated OLS estimated coefficients, working with the y_i 's and x_i 's, compare to what you get had you had access to the true data, the y_i^* 's and x_i^* 's

Measurement: Units and Errors

's? Or put differently, Can we say anything about the relationship between what we can estimate, $\hat{\beta}_1$, and what we would have estimated in the absence of measurement error, $\hat{\beta}_1^*$. Perhaps surprisingly, we can in fact say something about this in a small number of cases. In general however, everything depends on the specific nature of the errors.

- 35) Classic mismeasurement assumptions: In the classic analysis of errors in variables, which goes back to the late 1800's, we assume that the errors are uncorrelated with one another, and as well uncorrelated with the true values of the variables.³ To avoid trivial cases, we'll assume that the measurement errors have non-zero variances.
- 36) To start, and invoking the classic assumptions, let's look at three simple cases:
- a) Case I: Errors on the LHS
 - b) Case II: Errors on the RHS
 - c) Case III: Errors on Both Sides

Case I - Errors on the LHS: Measurement error with the y's only

- 37) In the first LHS Case, the x's have no measurement error, and the y's have measurement error, Δy . Further, and under the classic assumptions, the covariance of that measurement error with x is 0 (as would be the case if the measurement error with the y's was independent of the x's). So: $x = x^*$, $y \neq y^*$ and $S_{x\Delta y} = S_{x\Delta y} = 0$.

- 38) Since $\hat{\beta}_1^* = \frac{S_{y^*x}}{S_{x^*x}}$, and since $x = x^*$, any difference between $\hat{\beta}_1$ and $\hat{\beta}_1^*$ will be driven entirely by the extent to which S_{yx} differs from S_{y^*x} (the extent to which the sample covariance of y with x differs from the sample covariance of y^* with x).

39) But then
$$S_{yx} = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{n-1} = \frac{\sum [(y_i^* + \Delta y_i) - (\bar{y}^* + \Delta \bar{y})](x_i - \bar{x})}{n-1}$$
$$= \frac{\sum (y_i^* - \bar{y}^*)(x_i - \bar{x})}{n-1} + \frac{\sum (\Delta y_i - \Delta \bar{y})(x_i - \bar{x})}{n-1} = S_{y^*x} + S_{x\Delta y} = S_{y^*x} \text{ since } S_{x\Delta y} = 0.$$

- 40) And so in this case, $\hat{\beta}_1 = \hat{\beta}_1^*$.. or put in words: there will be no impact on the estimated slope coefficient if the mismeasurement of the y's is uncorrelated with the x's, and if there is no measurement error with the x's.⁴

³ Throughout this discussion, and because we are looking at the impact on OLS estimates, terms like correlation, covariance and variance will refer to the sample statistics for the given data... and not properties of random variables.

⁴ The intercept estimate will be impacted to the extent that the mean of the y errors differs from 0.

Measurement: Units and Errors

Case II - Errors on the RHS: Measurement error with the x's only

41) In Case II, the y's have no measurement error. The x's have measurement error, Δx , which has non-zero sample variance, and is uncorrelated with both the x's and the y's. So: $x \neq x^*$, $y = y^*$, $S_{\Delta x \Delta x} > 0$ and $S_{x^* \Delta x} = S_{y^* \Delta x} = S_{y \Delta x} = 0$.

$$42) \text{ In this case, } S_{yx} = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{n-1} = \frac{\sum (y_i^* - \bar{y}^*)[(x_i^* + \Delta x_i) - (\bar{x}^* + \bar{\Delta x})]}{n-1}$$

$$= \frac{\sum (y_i^* - \bar{y}^*)(x_i^* - \bar{x}^*)}{n-1} + \frac{\sum (y_i^* - \bar{y}^*)(\Delta x_i - \bar{\Delta x})}{n-1} = S_{y^* x^*} + S_{y^* \Delta x} = S_{y^* x^*} \text{ since } S_{y^* \Delta x} = 0.$$

$$43) \text{ However, } S_{xx} = S_{x^* x^*} + S_{\Delta x \Delta x}, \text{ since } S_{x^* \Delta x} = 0, \text{ and so } \hat{\beta}_1 = \frac{S_{y^* x^*}}{S_{x^* x^*} + S_{\Delta x \Delta x}} = \frac{S_{x^* x^*}}{S_{x^* x^*} + S_{\Delta x \Delta x}} \hat{\beta}_1^*.$$

Since $\frac{S_{x^* x^*}}{S_{x^* x^*} + S_{\Delta x \Delta x}} < 1$, $|\hat{\beta}_1| < |\hat{\beta}_1^*|$. And so in this Case, the impact of measurement error on the RHS is to reduce the magnitude of the estimated slope coefficient... you have underestimated the magnitude of the effect!

Case III - Errors on Both Sides: Measurement error with the x's and y's

44) Under the classic assumptions, $S_{x^* \Delta x} = S_{y^* \Delta x} = S_{x^* \Delta y} = S_{y^* \Delta y} = S_{\Delta y \Delta x} = 0$.

$$45) \text{ Since } S_{yx} = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{n-1} = \frac{\sum [(y_i^* + \Delta y_i) - (\bar{y}^* + \bar{\Delta y})][(x_i^* + \Delta x_i) - (\bar{x}^* + \bar{\Delta x})]}{n-1}$$

$$= S_{y^* x^*} + S_{y^* \Delta x} + S_{x^* \Delta y} + S_{\Delta y \Delta x} = S_{y^* x^*}, \text{ and since } S_{xx} = S_{x^* x^*} + S_{\Delta x \Delta x}, \text{ we have the same result as in Case II, namely: } \hat{\beta}_1 = \frac{S_{y^* x^*}}{S_{x^* x^*} + S_{\Delta x \Delta x}} = \frac{S_{x^* x^*}}{S_{x^* x^*} + S_{\Delta x \Delta x}} \hat{\beta}_1^*, \text{ and } |\hat{\beta}_1| < |\hat{\beta}_1^*|, \text{ since } \frac{S_{x^* x^*}}{S_{x^* x^*} + S_{\Delta x \Delta x}} < 1.$$

46) And so under the classic mismeasurement assumptions, we have:

- a) Case I - Errors on the LHS: No impact; $\hat{\beta}_1 = \hat{\beta}_1^*$
- b) Case II: Errors on the RHS: Impact; $\hat{\beta}_1 = \frac{S_{x^* x^*}}{S_{x^* x^*} + S_{\Delta x \Delta x}} \hat{\beta}_1^*$ and $|\hat{\beta}_1| < |\hat{\beta}_1^*|$
- c) Case III: Errors on Both Sides: Impact; $\hat{\beta}_1 = \frac{S_{x^* x^*}}{S_{x^* x^*} + S_{\Delta x \Delta x}} \hat{\beta}_1^*$ and $|\hat{\beta}_1| < |\hat{\beta}_1^*|$

Measurement: *Units and Errors*

More Generally...

- 47) Unfortunately, once we move beyond the classic assumptions, the world can get very complicated... few general rules apply and the impact of mismeasurement needs to be evaluated on a case-by-case basis.
- 48) **My bias:** It's easy for econometricians to get very excited about errors in variables models. And no doubt, measurement error can be an issue. But those errors are almost always totally swamped by potential *endogeneity* issues.
- a) So while you should know about the possibility of errors in variables and be generally familiar with how those errors might impact your estimates, you should get back to work and focus on omitted variable bias... and ***Build a better model!***