PPOL502-01, Spring 2016
Course Notes #2: Data Scaling; Log Transformations; Absolute vs. Relative Change Interpretations

---

I. DATA SCALING

- Wooldridge covers this topic on pp. 40-41, and again in chapter 6 (pp. 186-189).

- In addition to reviewing those sections, the next few pages of these notes provide some additional examples of data scaling and its effects (or lack of effects).

```
. summarize wage wagedol lwage exper exper10
```

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        wage |      3010    577.2824    262.9583         100       2404
     wagedol |      3010    5.772824    2.629583           1      24.04
       lwage |      3010    6.261832    .4437976     4.60517    7.784889
       exper |      3010    8.856146    4.141672           0         23
     exper10 |      3010    .8856146    .4141672           0        2.3
```

```
. pwcorr wage wagedol lwage exper exper10, sig
```

```
             |     wage  wagedol    lwage    exper  exper10
-------------+-----------------------------------------------
        wage |   1.0000
             |
             |
     wagedol |   1.0000   1.0000
             |   0.0000
             |
       lwage |   0.9472   0.9472   1.0000
             |   0.0000   0.0000
             |
       exper |   0.0300   0.0300   0.0125   1.0000
             |   0.0995   0.0995   0.4936
             |
     exper10 |   0.0300   0.0300   0.0125   1.0000   1.0000
             |   0.0995   0.0995   0.4936   0.0000
```

Dependent Variable: wage

```
. regress wage exper
```

```
      Source |       SS       df       MS              Number of obs =    3010
-------------+------------------------------           F(  1,  3008) =    2.72
       Model |  187693.228     1  187693.228           Prob > F      =  0.0995
    Residual |   207875837  3008  69107.6585           R-squared     =  0.0009
-------------+------------------------------           Adj R-squared =  0.0006
       Total |   208063530  3009  69147.0688           Root MSE      =  262.88
------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   1.906942   1.157115     1.65   0.099    -.3618734    4.175758
       _cons |   560.3942   11.31248    49.54   0.000     538.2133    582.5752
------------------------------------------------------------------------------
```

Dependent Variable: wagedol

. regress wagedol exper

Source	SS	df	MS	Number of obs	3010
Model	18.7693241	1	18.7693241	F(1, 3008)	2.72
Residual	20787.5837	3008	6.91076587	Prob > F	0.0995
Total	20806.3531	3009	6.9147069	R-squared	0.0009
				Adj R-squared	0.0006
				Root MSE	2.6288

wagedol	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
exper	.0190694	.0115711	1.65	0.099	-.0036187	.0417576
_cons	5.603942	.1131248	49.54	0.000	5.382133	5.825752

Dependent Variable: wage

. regress wage exper10

Source	SS	df	MS	Number of obs	3010
Model	187693.228	1	187693.228	F(1, 3008)	2.72
Residual	207875837	3008	69107.6585	Prob > F	0.0995
Total	208063530	3009	69147.0688	R-squared	0.0009
				Adj R-squared	0.0006
				Root MSE	262.88

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
exper10	19.06942	11.57115	1.65	0.099	-3.618734	41.75758
_cons	560.3942	11.31248	49.54	0.000	538.2133	582.5752

Dependent Variable: wagedol

. regress wagedol exper10

Source	SS	df	MS	Number of obs	3010
Model	18.7693241	1	18.7693241	F(1, 3008)	2.72
Residual	20787.5837	3008	6.91076587	Prob > F	0.0995
Total	20806.3531	3009	6.9147069	R-squared	0.0009
				Adj R-squared	0.0006
				Root MSE	2.6288

wagedol	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
exper10	.1906942	.1157115	1.65	0.099	-.0361873	.4175758
_cons	5.603942	.1131248	49.54	0.000	5.382133	5.825752

Dependent Variable: lwage

. regress lwage exper

Source	SS	df	MS	Number of obs	3010
Model	.09236512	1	.09236512	F(1, 3008)	0.47
Residual	592.549246	3008	.196991106	Prob > F	0.4936
Total	592.641611	3009	.196956335	R-squared	0.0002
				Adj R-squared	-0.0002
				Root MSE	.44384

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
exper	.0013377	.0019536	0.68	0.494	-.0024928	.0051683
_cons	6.249985	.0190993	327.24	0.000	6.212536	6.287434

II. LOGARITHMIC FUNCTIONAL FORMS

[note: these notes draw heavily on Mirer 1995, *Economic Statistics and Econometrics* 3rd ed.]

- Wooldridge introduces logarithmic functional forms in chapter 2 (pp. 41-44), and again discusses them in chapter 6 (pp. 191-194).

- Step back: remember OLS estimates models that are LINEAR in the parameters. So, for example, we can include quadratic or cubic X terms in an equation (we'll cover this in chapter 6) – the parameter estimate itself is still a linear relationship.

- The basic idea with logarithms is this: some nonlinear relationship become linear (and thus is suitable for estimation by OLS) when it is transformed with a logarithm.

- What ARE logarithms? They are basically a transformation of data (see handout on Napier for why they were originally useful). Think back to your math classes when you dealt with *e* (which here indicates the base of natural logarithms, not the error from a regression), natural logs, and rules of logs (also see Wooldridge Appendix A. Wooldridge's notes on the natural log are on pp. 712-716 if you want extra info]

- It is common in practice to use the log transformation when the variable is a positive dollar amount (e.g., wages, salaries, firm sales) or characterized by large integer values (e.g., population, total number of employees, school enrollment).

- The following few pages provide an overview for interpreting coefficients from regression models with logged variables. The graphs from Mirer are very helpful for getting a picture of these functions. Also, supplementary materials at the end of these notes are intended to give you a better sense of logarithmic functions.

** This table from Wooldridge's p. 44 provides helpful reminders for how to interpret log models:

SUMMARY TABLE FROM WOOLDRIDGE

Model	**Dependent Variable**	**Independent Variable**	**Interpretation of β_1**
level-level	y	x	$\Delta y = \beta_1 \Delta x$
level-log (uncommon)	y	log(x)	$\Delta y = (\beta_1/100)\%\Delta x$
log-level (semi-elasticity of y vs x)	log(y)	x	$\%\Delta y = (100\beta_1)\Delta x$ (constant % effect)
log-log (elasticity of y vs x)	log(y)	log(x)	$\%\Delta y = \beta_1 \%\Delta x$ (constant elasticity)

A. LOG-LOG SPECIFICATIONS

For example, suppose the following describes the relationship between X and Y:

$$Y = (e^{\beta_0})(X^{\beta_1})$$

Take the natural log of both sides of the equation, and obtain:

$$\ln(Y) = \beta_0 + \beta_1 \ln(X)$$

This expression is linear between X and Y and is known as a *log-linear relationship* between X and Y.

[note: Wooldridge uses the notation: $\log(Y) = \beta_0 + \beta_1 \log(X)$, where he assumes that "log" is the natural log "ln". The "log" function in STATA also computes the natural log.)

- What does this relationship look like? It depends on the value of β_1. **[see handout from Mirer, p. 120]**

- Another important thing to note is that β_1 is the *point elasticity* of Y with respect to X, which is constant over the whole range of the relation, and so is also known as a *constant elasticity model*.

$\beta_1 = \dfrac{dY/Y}{dX/X}$, where dY and dX can be thought of as very small changes in Y and X.

β_1 has a slightly different interpretation in this model compared to the typical interpretation of the slope coefficient that we talked about for a "regular" X (think *percentages)*

Example:

$$\ln(\text{MONEY SUPPLY}) = \beta_0 + \beta_1 \ln(\text{GNP}) + u$$

$$\ln(\text{MONEY SUPPLY}) = 3.948 + 0.215 \ln(\text{GNP})$$

- Question: With this specification, what does the relationship between MS and GNP look like (refer back to handout from Mirer)?

- Interpretation:

  (a) The elasticity of MS with respect to GNP is 0.215. If GNP increases by 1 percent, we predict that money supply will increase by 0.215 percent.

  (b) If GNP increase by 5 percent, we predict that money supply will increase by (0.215)*(5) = 1.075 percent.

  (c) When the changes to GNP or MS are large, the approximation above isn't very good (remember, we are estimating a *point* elasticity with this model).

  In these large-change cases, we need to calculate the exact change expected, using the following:

  proportionate change of Y = $\exp[\beta_1 * \ln(1 + \text{proportionate change of X})] - 1$

  So, if GNP were to increase by 100 percent (i.e., proportionate change of X = 1.00):

  proportionate change of Y = exp[0.215*ln (1 + 1)] -1

  = exp[0.215*0.693] -1
  = 1.16 – 1 = 0.16

  So, a 100 percent increase in GNP is predicted to lead to a 16 percent increase in money supply, using this exact formula. What would we have predicted with the method in (b)? (0.215)*(100) = 21.5% percent increase in money supply.

B. SEMI-LOG SPECIFICATIONS

For example, suppose you have the following relation between X and Y: $Y = (e^{\beta_0})(e^{\beta_1 X})$

If you take the natural log of both sides of the equation, you obtain: $\ln(Y) = \beta_0 + \beta_1 X$

This expression is linear between X and Y and is known as a *semi-log relationship* between X and Y.

- What does this relationship look like? It depends on the value of β_1. **[see handout from Mirer, p. 123]**

- Now,

$\beta_1 = \dfrac{dY/Y}{dX}$, where dY and dX can be thought of as very small changes in Y and X.

β_1 has a slightly different interpretation in this model: It is the ratio of the proportionate change in Y to the absolute change in X.

Example:

$\ln(EARNS) = \beta_0 + \beta_1 EDUC + u$

$\ln(EARNS) = 0.673 + 0.107\, EDUC$

- Question: What does the relationship between EDUC and EARNS look like?

- Interpretation:

(a) An additional year of schooling (EDUC) is predicted to increase earnings by the proportion 0.107, or 10.7 percent.

(b) If schooling increases by 3 years, we predict that earnings will increase by (0.107)*(3) = 0.321, or 32.1 percent.

(c) When the changes to EDUC or EARNINGS are large, the approximation above isn't very good.

In these large-change cases, we need to calculate the exact change, using the following formula:

proportionate change of Y = exp[β_1 *change in X] – 1

So, if schooling were to increase by 3 years:

proportionate change of Y = exp[0.107*3]- 1

= exp[0.321]-1
= 1.379-1 = 0.379, or 37.9 percent

C. THREE EXAMPLES: LEVEL-LEVEL, LOG-LEVEL, LOG-LOG

1. For a unit increase in x, y increases by constant absolute amount (50)

x	y	ln(y)
1	100	4.60517
2	150	5.010635
3	200	5.298317
4	250	5.521461
5	300	5.703782
6	350	5.857933
7	400	5.991465
8	450	6.109248
9	500	6.214608
10	550	6.309918

Correlation Matrix

	x	*y*	*ln(y)*
x	1		
y	1	1	
ln(y)	0.972575	0.972575	1

2. For a unit increase in x, y increases by constant percent (50%)

x	y	ln(y)
1	100	4.60517
2	150	5.010635
3	225	5.4161
4	337.5	5.821566
5	506.25	6.227031
6	759.375	6.632496
7	1139.063	7.037961
8	1708.594	7.443426
9	2562.891	7.848891
10	3844.336	8.254356

Correlation Matrix

	x	*y*	*ln(y)*
x	1		
y	0.900585	1	
ln(y)	1	0.900585	1

3. For a relative increase in x (e.g. 10%), y increases by constant relative amount (e.g. 50%)

x	ln(x)	y	ln(y)
1.00	0	100	4.60517
1.10	0.09531	150	5.010635
1.21	0.19062	225	5.4161
1.33	0.285931	337.5	5.821566
1.46	0.381241	506.25	6.227031
1.61	0.476551	759.375	6.632496
1.77	0.571861	1139.063	7.037961
1.95	0.667171	1708.594	7.443426
2.14	0.762481	2562.891	7.848891
2.36	0.857792	3844.336	8.254356

Correlation Matrix

	x	*y*	*ln(x)*	*ln(y)*
x	1			
y	0.944637	1		
ln(x)	0.992863	0.900585	1	
ln(y)	0.992863	0.900585	1	1

LOGARITHMS USING EXCEL:

LN

Returns the natural logarithm of a number. Natural logarithms are based on the constant e (2.71828182845904).

LN(number)

Number is the positive real number for which you want the natural logarithm.

Remark

LN is the inverse of the EXP function.

	A	B
1	**Formula**	**Description (Result)**
2	=LN(86)	Natural logarithm of 86 (4.454347)
3	=LN(2.7182818)	Natural logarithm of the value of the constant e (1)
4	=LN(EXP(3))	Natural logarithm of e raised to the power of 3 (3)

EXP

Returns e raised to the power of number. The constant e equals 2.71828182845904, the base of the natural logarithm.

EXP(number)

Number is the exponent applied to the base e.

Remarks

To calculate powers of other bases, use the exponentiation operator (^).
EXP is the inverse of LN, the natural logarithm of number.

	A	B
1	**Formula**	**Description (Result)**
2	=EXP(1)	Approximate value of e (2.718282)
3	=EXP(2)	Base of the natural logarithm e raised to the power of 2 (7.389056)

LOG

Returns the logarithm of a number to the base you specify.

Syntax

LOG(**number**,base)

Number is the positive real number for which you want the logarithm.

Base is the base of the logarithm. If base is omitted, it is assumed to be 10.

	A	B
1	**Formula**	**Description (Result)**
2	=LOG(10)	Logarithm of 10 (1)
3	=LOG(8, 2)	Logarithm of 8 with base 2 (3)
4	=LOG(86, 2.7182818)	Logarithm of 86 with base e (4.454347)

D. ABSOLUTE VS. PROPORTIONATE VS. PERCENTAGE VS. PERCENTAGE POINT CHANGES (also see Wooldridge Appendix A.3)

Proportionate and percentage changes for things like dollar amounts or population sizes (interval-ratio variables) are fairly straightforward:

- Example 1: Income (Interval Ratio Variable)

  If an individual's annual income goes from $30,000 in 1994 to $36,000 in 1995, then there was an *absolute change* of $6,000. However, the *proportionate change* in income is 6,000/30,000=.20. Multiplying this proportionate change of .20 by 100 converts to a *percentage change* of 20%.

  Note that it makes no sense to talk of a *percentage point* change in income because it is not measured as a percentage (Example 2 below gives an example where a percentage point change has meaning).

Interpreting percentage changes can be tricky when the variable of interest itself is a percentage:

- Example 2: Completed College Education (No, Yes)

  Suppose that the percentage of adults with a college education in a particular city was 24% in 1990 and 30% in 2000. The absolute change from 1990 to 2000 is 6 *percentage points*, <u>not</u> 6%!!! The *proportionate change* in the percentage of the city's population with a college education is 6/24=.25, which can be converted to a *percentage change* of 25%.

  Note: The phrases *percentage change* or *percent change* are synonymous.

Now let's look at an example of interpreting coefficients from two different simple linear regression models (level-level and log-level) where both the dependent and independent variables are percentages.

- Example 3: Simple Linear Regression Model

The Wooldridge MEAP93 data set contains summary data for a sample of 408 schools. Suppose we are interested in looking at the relationship between the percent of students passing MEAP science (Y) and the percent passing MEAP math (X). Descriptive stats for the dependent (sci11) and independent (math10) variables and the correlation between them was obtained using pwcorr in Stata (see code and output on next page):

```
. summ sci11 math10

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       sci11 |       408    49.18309    12.52467        7.2       85.7
      math10 |       408    24.10686    10.49361        1.9       66.7


. pwcorr sci11 math10, obs sig

             |    sci11   math10
-------------+------------------
       sci11 |   1.0000
             |
             |      408
             |
      math10 |   0.2261   1.0000
             |   0.0000
             |      408      408
```

MODEL 1 (level-level): Regress sci11 on math10

```
. regress sci11 math10

      Source |       SS       df       MS              Number of obs =     408
-------------+------------------------------           F(  1,   406) =   21.87
       Model |  3263.33861     1  3263.33861           Prob > F      =  0.0000
    Residual |  60581.6534   406  149.215895           R-squared     =  0.0511
-------------+------------------------------           Adj R-squared =  0.0488
       Total |   63844.992   407  156.867302           Root MSE      =  12.215

------------------------------------------------------------------------------
       sci11 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      math10 |   .2698415   .0577013     4.68   0.000      .156411    .3832721
       _cons |   42.67806   1.516772    28.14   0.000     39.69635    45.65976
------------------------------------------------------------------------------
```

Interpretations for *math10* coefficient:

- For a 1 *percentage point* increase in the percent of students passing MEAP math, we predict that the percent of students passing MEAP science increases by 0.27 *percentage points.*

- For a 10 *percentage point* increase in the percent of students passing MEAP math, we predict that the percent of students passing MEAP science increases by 2.7 *percentage points.*

MODEL 2 (refer to as log-level or semi-log): Regress lnsci11* on math10

```
. gen lnsci11=ln(sci11)

. regress lnsci11 math10
```

```
      Source |       SS       df       MS              Number of obs =     408
-------------+------------------------------           F(  1,   406) =   24.07
       Model |  2.18410907     1  2.18410907           Prob > F      =  0.0000
    Residual |  36.8397616   406  .090738329           R-squared     =  0.0560
-------------+------------------------------           Adj R-squared =  0.0536
       Total |  39.0238707   407  .095881746           Root MSE      =  .30123

------------------------------------------------------------------------------
     lnsci11 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      math10 |    .006981   .0014229     4.91   0.000     .0041838    .0097781
       _cons |   3.686979   .0374031    98.57   0.000     3.613451    3.760507
------------------------------------------------------------------------------
```

Examples of interpretations for *math10* coefficient:

- For a 1 *percentage point* increase in the percent of students passing MEAP math, we predict that the percentage of students passing MEAP science increases by about 0.7 *percent.*

- For a 10 *percentage point* increase in the percent of students passing MEAP math, we predict that the percentage of students passing MEAP science increases by about 7 *percent.*