New York University

Wilf Family Department of Politics

Prof. Patrick Egan

**QUANT I: Transforming X to Specify Diminishing Returns to X**

set obs 1000

egen x = fill(.01,.02)

gen ln\_x = ln(x)

egen y = fill (0,1)

gen ln\_y = ln(y)

twoway (line ln\_y y) (line ln\_x x), legend(off) xtitle("x") ytitle (ln(x))

\*A log transformation stretches out low values and compresses high values

\*thus when x’s marginal effect on y diminishes as x grows large (which is true of many social phenomena), the relationship between y and ln(x) is more likely to be linear than the relationship between y and x.

\*another interpretation: change in y is predicted less well by the absolute change in x than the percentage change in x.

\*Why does this matter? OLS Assumption 1: we assume we have correctly specified a population model that is linear in its parameters. Violation of this assumption can lead to bias and imprecision.

\* An example where we look at the relationships among crime, poverty and population density:

use "counties.dta", clear

twoway (scatter crimerate density, msize(small))

****

\*looks like there are some outliers. Let’s see what they are:

. list county state density if density>=10000

+---------------------------------------------+

| county state density |

|---------------------------------------------|

2245. | Yukon-Koyukuk AK . |

2295. | Yellowstone National Park MT . |

2296. | South Boston VA . |

2315. | Suffolk MA 11692 |

2806. | Queens NY 20453 |

|---------------------------------------------|

2808. | Hudson NJ 12957 |

2810. | New York NY 66835 |

2811. | Bronx NY 31730 |

2812. | Kings NY 34723 |

2867. | Philadelphia PA 11241 |

|---------------------------------------------|

2990. | San Francisco CA 16526 |

+---------------------------------------------+

\*note this includes all with value “.” which Stata treats as a very large number.

\*we don’t want these observations to play too strong a role in our analysis; leave them out for now.

twoway (scatter crimerate density, msize(tiny)) (lowess crimerate density, lw(thick)) if density<10000, legend(off)

graph save Graph "density.gph", replace

\*we have reason to think that while density is associated with crime, this association might be diminishing in density. To see if this is the case, let’s create a log transformation of density:

gen ln\_density = ln(density)

twoway (scatter crimerate ln\_density, msize(tiny)) (lowess crimerate ln\_density, lw(thick)) if density<10000, legend(off)

graph save Graph "ln\_density.gph", replace

graph combine "density.gph" "ln\_density.gph", ycommon

\* ln\_density appears to be a better predictor of crime than density.

\* Bivariate regressions confirm this:

. reg crimerate density if density<10000

Source | SS df MS Number of obs = 2662

-------------+------------------------------ F( 1, 2660) = 223.48

Model | 618563579 1 618563579 Prob > F = 0.0000

Residual | 7.3627e+09 2660 2767930.03 R-squared = 0.0775

-------------+------------------------------ Adj R-squared = 0.0772

Total | 7.9813e+09 2661 2999345.15 Root MSE = 1663.7

------------------------------------------------------------------------------

crimerate | Coef. Std. Err. t P>|t| [95% Conf. Interval]

-------------+----------------------------------------------------------------

density | .8876232 .0593764 14.95 0.000 .7711945 1.004052

\_cons | 2686.712 34.18831 78.59 0.000 2619.674 2753.75

------------------------------------------------------------------------------

. reg crimerate ln\_density if density<10000

Source | SS df MS Number of obs = 2662

-------------+------------------------------ F( 1, 2660) = 660.40

Model | 1.5874e+09 1 1.5874e+09 Prob > F = 0.0000

Residual | 6.3939e+09 2660 2403706.28 R-squared = 0.1989

-------------+------------------------------ Adj R-squared = 0.1986

Total | 7.9813e+09 2661 2999345.15 Root MSE = 1550.4

------------------------------------------------------------------------------

crimerate | Coef. Std. Err. t P>|t| [95% Conf. Interval]

-------------+----------------------------------------------------------------

ln\_density | 459.948 17.89808 25.70 0.000 424.8525 495.0436

\_cons | 1096.143 74.80349 14.65 0.000 949.4638 1242.822

------------------------------------------------------------------------------

\*interpretation: a one-percent increase in population density is associated with an increase in the crime rate of approximately 460 crimes per 100,000 people.

\*This improved specification has implications for our estimates of other variables. Let’s say we want to look at the association between poverty and crime, controlling for density. First look at correlation matrix:

. pwcorr povrate crimerate density ln\_density if density<10000

| povrate crimer~e density ln\_den~y

-------------+------------------------------------

povrate | 1.0000

crimerate | 0.0643 1.0000

density | -0.1174 0.2784 1.0000

ln\_density | -0.2511 0.4460 0.5579 1.0000

\*ln\_density is much more highly correlated with poverty and crime than density. This will have implications for our estimates.

. reg crimerate povrate density if density<10000

Source | SS df MS Number of obs = 2662

-------------+------------------------------ F( 2, 2659) = 128.38

Model | 702809198 2 351404599 Prob > F = 0.0000

Residual | 7.2784e+09 2659 2737287.8 R-squared = 0.0881

-------------+------------------------------ Adj R-squared = 0.0874

Total | 7.9813e+09 2661 2999345.15 Root MSE = 1654.5

------------------------------------------------------------------------------

crimerate | Coef. Std. Err. t P>|t| [95% Conf. Interval]

-------------+----------------------------------------------------------------

povrate | 28.90484 5.210233 5.55 0.000 18.68832 39.12136

density | .9321349 .0595895 15.64 0.000 .8152884 1.048981

\_cons | 2246.886 86.26309 26.05 0.000 2077.737 2416.036

------------------------------------------------------------------------------

povrate estimated more precisely because effect of density is now more properly specified.

. reg crimerate povrate ln\_density if density<10000

Source | SS df MS Number of obs = 2662

-------------+------------------------------ F( 2, 2659) = 414.21

Model | 1.8959e+09 2 947949932 Prob > F = 0.0000

Residual | 6.0854e+09 2659 2288588.79 R-squared = 0.2375

-------------+------------------------------ Adj R-squared = 0.2370

Total | 7.9813e+09 2661 2999345.15 Root MSE = 1512.8

------------------------------------------------------------------------------

crimerate | Coef. Std. Err. t P>|t| [95% Conf. Interval]

-------------+----------------------------------------------------------------

povrate | 57.07723 4.916074 11.61 0.000 47.43751 66.71694

ln\_density | 518.8822 18.18696 28.53 0.000 483.2202 554.5442

\_cons | 18.89011 118.0527 0.16 0.873 -212.5944 250.3746

------------------------------------------------------------------------------

Association between povrate and crimerate now found to be much larger. Looking back at the correlation matrix, can you see why?