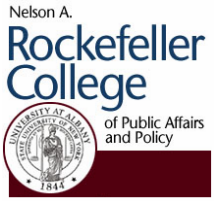

Rockefeller College
University at Albany

PAD 705 Handout: Standardized Coefficients

In some cases, knowing whether a coefficient is statistically significant or not exhausts our interest in a particular regression model. However, in many policy applications what we would really like to know is what the tradeoffs are. If both improving the pay of teachers and increasing the number of teachers improves test scores for kids, which should we do? Do I get more bang for the buck if I invest $100 in teachers pay or $100 in hiring more teachers? If I want to be admitted to a good school, what is the trade-off between a high GPA and high SAT scores? If I have a lousy GPA, will maxing my SAT score fill the gap? Is it more important to raise family income or lower the unemployment rate if I wish to lower the crime rate per 100,000 inhabitants? Should I invest in more jobs – maybe through a jobs program – or arrange for everyone with a job to get a raise – possibly through a deduction in the payroll tax? These are all questions that we might be able to answer from the coefficient estimates, but not the "raw", unstandardized ones that Stata generally reports. Instead, we need to have Stata produced standardized coefficients.

Why standardized coefficients?

Why can't we directly compare coefficients? Recall that the size of the coefficient depends in part on the mean and variance of the independent variable. Often, our independent variables have widely varying means and variances, meaning that the resulting coefficients can't be directly compared. Think about the GPA vs. SAT example. GPAs range from 0.0 to 4.0, with a mean that is somewhere around 2.4. SAT scores range from 400 to 1600, with a mean score of 1020 (in 2002-2003). Because the range is so much smaller for GPA, a 1 point increase in GPA would cause a huge increase in the probability of college admission, while a 1 point increase in SAT score would be associated with a very small increase. For this reason, the unstandardized coefficients Stata reports are not directly comparable. This is not always true; sometimes coefficients *are* directly comparable. For instance, if you wish to see if an additional year of education is more valuable than an additional year of experience, one need only compare the coefficients. But in many cases we need to standardize the coefficients in order to make comparisons.

How do we standardize?

What we really want to do is account for differences in the range and variance of the independent variables. The usual way this is done in statistics (not just in regression) is to convert the values of the independent variable into *deviations*. Deviations are very similar to a Z score. The idea is to express values for all independent variables in a dataset in terms of the number of standard deviations that variable is from the its mean. Procedurally, for each variable we should first calculate the mean and variance across all observations. Then for each observation, we would subtract out the mean and divide by the variance for each variable. So, if our regression equation is:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_3 i + \varepsilon_i$$

the equation used to find standardized regression coefficients (which I will denote as β^*) is:

$$\frac{y_i - \bar{y}}{s_y} = \beta^*_i \frac{x_{1i} - \bar{x}_1}{s_{x_1}} + \beta^*_2 \frac{x_{2i} - \bar{x}_2}{s_{x_2}} + \beta^*_3 \frac{x_{3i} - \bar{x}_3}{s_{x_3}} + \varepsilon_i$$

(Remember: there is no constant here – β_0. Why? See the footnote at the bottom if you can't figure it out.[*])

It turns out, that there is a simplier formula for converting from an unstandardized coefficient to a standardized one. Using β_1 as an example:

$$\beta^*_1 = \beta_1 \frac{s_{x_1}}{s_y}$$

The standardized coefficient is found by multiplying the unstandardized coefficient by the ratio of the standard deviations of the independent variable (here, x_1) and dependent variable. This makes some intuitive sense: the coefficient should be scaled to reflect the difference between the "spread" of the independent variable and the dependent variable.

The standardized coefficients then have a slightly different meaning. If $\beta_1^* = 0.55$, then a 1 standard deviation change in x_1 results in a 0.55 standard deviation increase in the dependent variable. The β^*s are comparable because they all refer to a 1 standard deviation change in their respective independent variables rather than a one unit change.

Finding standardized coefficients in Stata

Let's take a look at another example drawn from data on crime rates per 100,000 inhabitants in a sample of metropolitan statistical areas (MSAs) from across the United States. The working hypothesis here is that (a) crime rates increase as the proportion of people who drop out of high school (popelpls) increases and as civilian labor force unemployment (uerate) increases and (b) crime rates decrease as the median family income (medhhinc) in the MSA increases. Below is the regression output:

[*] The constant comes from inserting a column of ones into the dataset. The mean of a variable consisting of only 1s is 1 and there is no variance. In a standardized framework, there is no constant.

```
. reg  crimrate perelpls  medhhinc uerate

      Source |       SS       df       MS              Number of obs =     933
-------------+------------------------------           F(  3,   929) =   67.14
       Model |  1.4021e+09     3   467352216           Prob > F      =  0.0000
    Residual |  6.4662e+09   929  6960369.09           R-squared     =  0.1782
-------------+------------------------------           Adj R-squared =  0.1755
       Total |  7.8682e+09   932  8442317.09           Root MSE      =  2638.3

------------------------------------------------------------------------------
    crimrate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    perelpls |  -1038.292   1039.263    -1.00   0.318    -3077.867    1001.283
    medhhinc |  -.2028265   .0223986    -9.06   0.000    -.2467842   -.1588689
      uerate |   15835.57   3716.514     4.26   0.000     8541.831     23129.3
       _cons |   10009.58   649.8157    15.40   0.000     8734.305    11284.86
------------------------------------------------------------------------------
```

Interestingly, it appears that while there is the expected positive correlation between drop out rates and crime, the relationship is not statistically significant. However, median family income and civilian labor force variables have the expected relationships and are statistically significant. Given this finding, which one should we work on first? First, let's look at the variance in these independent variables.

```
. summ medhhinc, det

                  Median hsehld income 1979
-------------------------------------------------------------
      Percentiles      Smallest
 1%         9807           7710
 5%        11548           8459
10%        12105           8503       Obs                 956
25%        14012           8583       Sum of Wgt.         956

50%      16361.5                      Mean           17691.15
                        Largest       Std. Dev.      5133.377
75%      20727.5          38542
90%        25023          39926       Variance       2.64e+07
95%        27375          41143       Skewness       1.090808
99%        32373          41973       Kurtosis       4.474534


. summ uerate, det

                           uerate
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .0221822       .0161039
 5%     .0292138       .0189778
10%     .0343246       .0191704       Obs                 956
25%     .0439752       .0197006       Sum of Wgt.         956

50%     .0593052                      Mean           .0642176
                        Largest       Std. Dev.      .0282957
75%     .0797357       .1900266
90%     .1001192       .1995771       Variance       .0008006
95%     .1117774         .21052       Skewness       1.450573
99%     .1577931       .2492523       Kurtosis         7.1173
```

As you can see, the range and variance of the two variables is considerably different. Thus, we cannot directly compare the coefficients. So we need to standardize. It turns out that Stata includes an option for the "`regress`" command that reports standardized coefficients in the part of the output table where you would normally find the 95% confidence interval. The option is invoked by putting "`, beta`" at the end of the `regress` command.

```
. reg  crimrate perelpls  medhhinc uerate, beta

      Source |       SS       df       MS              Number of obs =     933
-------------+------------------------------           F(  3,   929) =   67.14
       Model |  1.4021e+09     3   467352216           Prob > F      =  0.0000
    Residual |  6.4662e+09   929  6960369.09           R-squared     =  0.1782
-------------+------------------------------           Adj R-squared =  0.1755
       Total |  7.8682e+09   932  8442317.09           Root MSE      =  2638.3

------------------------------------------------------------------------------
    crimrate |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
    perelpls |  -1038.292   1039.263    -1.00   0.318                -.0415989
    medhhinc |  -.2028265   .0223986    -9.06   0.000                -.3552206
      uerate |   15835.57   3716.514     4.26   0.000                 .1553135
       _cons |   10009.58   649.8157    15.40   0.000                        .
------------------------------------------------------------------------------
```

Using the `beta` option, Stata reports both the unstandardized and standardized coefficients. Why? *Because you cannot run statistical significance tests against the standardized coefficients.* You must use the unstandardized coefficients because the standard errors are calculated with reference to them, not the standardized coefficients.

Take a look at the "Beta" column. If we increase median family income by 1 standard deviation, crime rates per 100,000 people will decrease by 0.355 standard deviations while decreasing the unemployment rate by 1 standard deviation will result in only a 0.155 standard deviations decrease. It appears that, in terms of standard units, increasing median household income is more than twice as effective as decreasing the unemployment rate. In fact, increasing median income is 0.355/0.155 = 2.29 times as effective.

As policymakers, what should we do? The answer requires more analysis and information. The answer depends on what it costs to create a one standard deviation change in both measures. Let's say that raising the median income one standard deviation in Doctorateland costs $10 million dollars in direct subsidies to low income families. Then the cost of lowering the unemployment rate in Doctorateland one standard deviation must be no more than $10 million/2.29 = $4.37 million, if we wish to get the same decrease in crime that the $10 million dollar investment in family subsidies would yield.