

Assignment 6

Due Monday November 6, 6pm. Hard copy (with an updated Project Description if revised from Assignment 5) should be handed in during class, put under Prof. Kahn's door (518C)
(Worth 1 point if completed and handed in on time; late assignments lose 5% per day)

1. Post your current Stata data set on the google drive, replacing your previous data set (and again titled with your name e.g. janedoe.dta) at: <https://tinyurl.com/qm222a1>
2. Make any revisions necessary to Project Description questions 1-8.
3. Run additional multiple regressions. Specifically:
 - a. Think hard about whether there are additional omitted variables (i.e. confounding factors) that you *can* measure that are likely to be biasing your key coefficient(s). If you can find data on them, add them into the regressions. (If you really cannot think of anything beyond what you have, just write that.)
 - b. Identify at least one omitted variable that you *cannot* measure, reason out the sign of the omitted variable bias and explain here in 1-3 sentences why and in what direction it will bias your key coefficient.
 - c. If you have any numeric explanatory (X) variable, add a quadratic term in addition to your other variables to test if this nonlinear specification better. (If you are good at math and prefer to add a different nonlinear variable or to make your dependent variable non-linear, be my guest.) **Copy and paste here** the related command line and Stata regression output from the Results panel. (If you do not have *any* numeric explanatory variable, just say that.)
 - d. Explain here what you learn from this result (1-3 sentences).
 - e. If you have a numeric explanatory (X) variable that is very skewed, think about whether top-coding or taking the log of that variables is appropriate instead of or in addition to the quadratic. Explain here why you believe that it is appropriate. Then run the new regression with this variable (along with the other X variables). **Copy and paste here** the related command line and Stata regression output from the Results panel. (If you do not have *any* numeric explanatory variable, just say that.)
 - f. Think about if you can and should use “interaction terms” which means different slopes for the same variable, depending on a different variable. (Like the Budget*SciFi example. This will be most useful if you think that different groups have different slopes.) Try at least one interaction term out in a multiple regression (with all your other variables as well). **Copy and paste here** the related command line and Stata regression output from the Results panel.
 - g. Explain here what you learn from this interaction term result (1-3 sentences).
 - h. Which is the best regression or set of regressions that you will use in your project? Why is it best? Choose either the regression in Question 3 part a, c, e, or f or run an additional regression that combines more than one of these special terms. (If you choose a new regression, **copy and paste here** the related command line and Stata regression output from the Results panel.)