

Homework Exercise 2
(due March 12 – Good luck!)

1. In the “countries 2002” data set, run the regression predicting a country’s life expectancy (average life expectancy for men and women) with its percent urban population and its status as a developing or developed country. Interpret all relevant coefficients and summary statistics, including the slope and intercept estimates for both developing and developed countries. Now estimate an alternative variable model predicting life expectancy with a country’s urban population, its status as a developing or developed country, and the *interaction* between urban population and developing/developed status. Interpret the intercept, slope coefficients and summary statistics from this alternative model. Test whether the slopes for urbanization for developed and for developing countries are each statistically different from zero. Graph the two regression lines from the alternative model with their respective confidence bands. Using all the information from this question, assess whether the additive or interactive model is superior, and what it means substantively about the effects of these variables on life expectancy.
2. Using the same data set, regress the life expectancy variable against the country’s number of doctors per 10,000 people, the country’s gross domestic product, and the interaction between doctors and gdp. Interpret the intercept, slope coefficients and summary statistics for this model. Graph the effect (regression coefficient) of doctors on life expectancy as gdp increases (with the confidence band included) and interpret the findings. Finally, discuss what all of the results tell you substantively about the effects of these variables on life expectancy.
3. Using the “bank-salaries” data, plot the relationship between age and current salary. Why would you expect to find a non-linear relationship between these two variables? Estimate a model that takes this non-linearity into account. Interpret the coefficients for the independent variables as well as the summary statistics for the model as a whole. Graph the predicted values of salary from the regression as age increases. Finally, calculate the point on X at which the relationship is estimated to change directions and interpret this result.
4. Using the same data set, consider the relationship between an employee’s current salary and his/her education level.
 - a. Would you suspect theoretically that there might be problems using OLS here because of heteroskedasticity? Why or why not? Run the OLS regression and examine the scatterplot of the residuals against X. Do they appear to be heteroskedastic?
 - b. Test statistically for non-constant error variance using a) the Goldfield-Quant test, comparing the residual variance for individuals who have 12 years or less of education with the residual variance of those with 15 years or more; and b) the White test. Interpret these test results.
 - c. Correct for the problem (if it exists) with weighted least squares and/or OLS with “heteroskedastic-consistent” standard errors. Which of these models is your preferred model, and why?

EXTRA CREDIT (5 points): From problem 1 above: assume that urbanization “mediates” the relationship between a country’s level of development and life expectancy. Discuss using the causal mediation framework, calculate relevant direct and indirect effects, and interpret the results. [HINT: The interpretations will be easier if you create a “DEVELOPED” variable where 0 is “developing” and 1 is “developed” and use that instead of “DEVELOPING”].