

Assignment #1

DUE: October 7th before class (1:00 PM)

Late assignments will be penalized at the rate of 10% per day

Total points: 100 possible

You may use either R or STATA to complete this assignment. Please submit a compiled version of your code that includes the output. In R, this can be achieved by “Compile Report” button in RStudio; in STATA this can be done by creating a .log file. On the first line of your code after the .log or has been started (if using STATA) please also include the following:

In R: `name <- Sys.info()
name[7]`
In STATA: `display "`c(username)'"`

Use “Dataset1,” available in the Assignments folder on Quercus, to carry out the analysis below. Unless indicated, a complete answer to each part of the assignment will include either a regression table or a figure—both should well-formatted and easily readable—and text describing and interpreting what you are presenting. If, for any part of the assignment observations are missing, those countries should be excluded (this should be obvious from the regression tables, which should always report sample size used in the regression).

Recall that when it comes to grading, commenting your code is your friend. It helps you organize your thoughts and communicates to me more about what you are trying to do.

We are interested in understanding country differences in the life expectancy at birth for females (LEBF). You can learn more about this data [here](#).

1. Draw a preliminary DAG suggesting a relationship between the following variables and LEBF: gross domestic product per capita (GDPPC), health expenditure per capita (HXPC), total fertility rate, and any other covariates in the data set you think are relevant. [10 points]
2. Do not forget to include relationships between the other independent variables in (1). What does your DAG tell you about interpreting any regression coefficients causally? [10 points] (Note that this part does not require a new figure or table.)
3. Make a table providing summary statistics for the variables listed in (1). Your table should include the mean, standard deviation, and sample size for each variable. Does anything of concern stand out to you? [10 points]
4. Regress LEBF on HXPC. Report the coefficients, standard errors, confidence intervals, p -values, R^2 , and sample size in a regression table. Interpret the table, noting the economic and statistical

significance of the relationship. What is the association between a 1,000-unit increase in GDPPC and LEBF? [10 points]

5. Now regress LEBF on HXPC and GDPPC. Discuss the results of this regression relative to those from (4). [10 points]
6. Do you recommend a nonlinear transformation for either GDPPC, HXPC, or LEBF? If so, defend your choice and repeat the regression in (5) with the appropriate transformations. Interpret how your results have changed. [10 points]
7. How might these results differ by geography? Create a variable that assigns each observation to a geographic region (e.g., continent) and report a regression that builds on (6) by including the appropriate dummy variables. Interpret your results. [10 points]
8. Finally, include an interaction term between HXPC and the indicator for African countries. What are you measuring with this interaction, and why might it be meaningful? Interpret the results of this coefficient. [10 points]
9. Why is establishing the causal relationship between GDDPC, HXPC, and LEBF difficult in a simple regression such as this? If possible, provide one key figure that highlights an identification problem in this scenario. [10 points]
10. What do your results from (9) and intuition suggest about the standard errors in your specification? Either justify the use of homoscedastic standard errors or implement a full specification with another, more robust method (e.g., heteroskedasticity-robust or clustered SEs). How does this change the results? [10 points]