HAD 5744: Health Econometrics 1
Dr. Hoagland
Acknowledgment: Rui Fu, Eric Nauenberg

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 using a "Knit" R-Markdown file with text and code; in STATA this can be done by creating a .log file. On the first line of your code (after the .log 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, 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. If you have coding questions, leave them in a comment on your code and I will try to provide feedback.

Question 1. 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), and total fertility rate. Include any other covariates in the data set you think are relevant (note that you do not have to include all of the variables in the data). Justify your DAG with text. [10 points]

2. Make sure that your DAG includes relationships between independent variables in (1), if needed. What does your DAG tell you about interpreting any regression coefficients (between LEBF, GDPPC, and HXPC) causally? [10 points]

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 HXPC and LEBF? (Note that HXPC is measured in dollars.) [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. (Note that there are multiple possible answers for this problem; the goal is to think critically about the causal identification.) [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]