**S-040: Introduction to Applied Data Analysis**

Harvard Graduate School of Education

**Homework 5: Including Dichotomous Predictors in Multiple Regression Models**

**DUE: Monday, December 2**

**ASSIGNMENT REQUIREMENTS**

You and your partner should submit 1 document for this assignment, including:

1. a cover page with your ID numbers and acknowledgments (IDs of those who helped you);
2. and a **7-page (double-spaced; this is not a typo, it’s just a shorter assignment and could be substantially shorter)** write-up of responses to questions 2-6.
3. your annotated code (will not count towards the page limit).

**Purpose of the assignment**

In this assignment, you will continue to use the dataset you used for Homework 4. In this assignment, however, we will be introducing a dichotomous variable, *HIGH\_VIOLENCE*, into the analysis. This variable identifies countries with high rates of violence, as measured by their murder rates; countries with murder rates above 5 murders per 100,000 people per year (chosen somewhat arbitrarily; we’re doing this because we need a dichotomous predictor) are identified as having high violence rates.

We’d like to expand our research question a little. Another important predictor of human development is the rate of violence within a country; countries experiencing high rates of violence tend to have lower levels of human development. However, violence might also be associated with educational spending; countries which are dealing with high rates of violence might not have as many resources to devote to education. We’d like our model to reflect this. We’re going to change our research question to be

**RQ: What is the relationship between educational spending and human development, adjusting for Gross National Income, rates of child dependency, and violence rates (i.e., identification of a country as high-violence)?**

**Dataset**

You will be using the same dataset that you used in Homework 5.

**1. Create a file to conduct the basic analysis**

Create a code file to conduct the statistical analyses necessary to complete this assignment. Annotate your code by providing a brief explanation of what each line of the program does (i.e., what action do you expect from each line of the do-file?). Use the code from your previous homework assignments and from lectures as guides, and feel free to reuse any code that is helpful to you.

Copy and paste your annotated do-file as an appendix. Be sure to copy the file in a monospace (i.e., fixed-width) font, such as Courier New, that is small enough that the spacing of your syntax and annotation looks like it does in your do-file window.

**<Submit code (in an appendix at the end of your document)>.**

**2. Create and describe the dichotomous predictor**

Create a new variable, high\_violence, which indicates whether a country has a high rate of violence, i.e., whether homicide\_rate is greater than or equal to 5.

In a very short paragraphdescribe the distribution of *HIGH\_VIOLENCE*. As part of your description, be sure to **describe what the mean of a dichotomous predictor** like *HIGH\_VIOLENCE* tells us.

**<Submit paragraph>.**

**3. Compare the mean human development of low- and high-violence countries**

Regress the outcome on *HIGH\_VIOLENCE* and begin constructing a taxonomy table that will present the results of this model. As in Homework 4, this table should be formatted in ‘presentation-style’, similar to the tables used in the lecture slides. Explain what we learn from **both fitted coefficients in this simple regression model**; make sure your explanations are substantive. Your answer should include **what you can say about this relationship in our sample** and **what can you say about this relationship in the population**.

**<Submit your answer>.**

**4. Examine the correlation matrix of the variables, and create table**

Add *HIGH\_VIOLENCE* to the correlation matrix you produced for Question 4 in the last assignment. Be sure that the table includes all of the appropriate correlations and their *p*-values. Write a paragraph that explains **what new information you learn from this correlation matrix** (specifically, how does *HIGH\_VIOLENCE* correlate with the outcome as well as the predictors we used in Homework 4?). Be sure to interpret your results in a substantive way. This could be a single very long sentence.

**<Submit paragraph and table>.**

**5. Does the relationship between the outcome and *HIGH\_VIOLENCE* changewhen you control statistically for the other variables?**

Go back to your final regression model from Homework 4. Add the results of this model to the taxonomy table you began to construct in Question 3 of this assignment. Now fit a model that includes the continuous predictors as well as the dichotomous predictor.Add this model to the taxonomy table. At this point the table should have **three** different models.

In a paragraph, compare the results of fitting these three models. In one or two sentences, describe what has changed and what has remained the same about the **estimate, standard error, and *p*-value** for the **coefficient of *L2EDUC***between the second and third models in your taxonomy. In another one or two sentences, describe what has changed and what has remained the same about the **coefficient of *HIGH\_VIOLENCE*’s estimate, standard error, and *p*-value** between the first and third models in your taxonomy. **At the end of this paragraph, please interpret the coefficient of *HIGH\_VIOLENCE* in your third model** (including what you can say about this relationship **in the sample** and what you can say about this relationship **in the population**), as well as the **95% confidence interval for the coefficient of *HIGH\_VIOLENCE***.

As part of your response to this question, submit the taxonomy table that you began constructing for Question 3 (this table should now include the results from three fitted regression models).

**<Submit paragraph and taxonomy table>.**

**6. Graphically display the results from your chosen model**

Create a **single display** that you believe is the **best visual representation** of the results of this analysis. Make sure that your plot shows at least two different fitted lines. We recognize that there are many viable alternatives; in choosing among them, be sure to consider the substantive points you want to make and create a plot that best allows you to highlight these conclusions. Write a brief paragraph explaining why you chose to construct your plot the way you did. **Make sure that your plot represents the controlled difference in predicted human development between low- and high-violence countries**.

**<Submit paragraph and figure>.**

**7. Correct a misinterpretation**

A reader is concerned that about your results. He claims that the reason that high-violence countries have lower levels of human development is because high violence countries are poorer, and less wealthy countries have lower levels of human development; he suggests that the difference you report between low- and high-income countries in your final model is simply reflecting that well-known fact. Drawing on the models you fit, **address the reader’s concern**. Make sure you reference **both the first and the final models in your table**. Be sure to **identify any truths to what the reader says**.

**<Submit your response (no longer than a page, and possibly less)>.**

**A final word**

If you have worked with other students in thinking through this assignment (remember, all writing *must* be done between you and your partner only), we ask you to please provide an explicit “authors’ acknowledgement” on the cover page of your submission in which you identify contributions to your work from anyone else other than your partner or the teaching team.