**Introduction to Statistics: Homework 1 (Multivariate Regression and Causality)**

**DUE THURSDAY, NOVEMBER 4th**

*Please type your responses. For the analysis in section 2 copy your regression output and paste it into your document. You can draw the causal diagrams by hand on a separate piece of paper and attach them to the back of your types answer document.*

1. One of the difficulties political researchers often have is sorting out how to measure a concept of interest. Imagine you wanted to examine the relationship between individuals' level of *generosity* and support for government-funded aid to the poor. You have a measure of support for government aid to the poor that ranges from 1 (strongly oppose) to 5 (strongly support).
   a. Write two survey questions that you could use to measure people's *generosity*. Include details about what the response options would be and their numerical ranges. [4 points]
   b. Make up some simple bivariate regression results (use the table below as a guide) predicting support for government aid to the poor with ONE of your two survey measures. You are making this up so it does not need to be "right" in the sense of what the real relationship looks like in the world. HOWEVER, it does need to be internally consistent (e.g., the T-values should be correct given the coefficient and standard errors you make up). [4 points]

| | Coefficient | Standard Error | T |
|---|---|---|---|
| [Your Variable] | ? | ? | ? |
| Constant | ? | ? | ? |

   c. Write out the regression equation (that you would use to calculate predicted values of support for government aid to the poor) based on this table. [4 points]
   d. What does the coefficient on the Constant mean? [4 points]
   e. What does the coefficient on [Your Variable] mean? [4 points]
   f. Think of one other variable that we might expect to be associated with support for government aid to the poor. Would adding this variable as an additional explanatory variable affect the coefficient on [Your Variable]? If so, how and why? If not, why not? [6 points]

2. For this set of questions use the “mortality” dataset. This data is from 1992. We are interested in what affects the level of infant mortality in a country.

   The variables in the dataset are:

   MORTINFT: Number of infant deaths per 1,000 live births.

   TVPERCAP: Number of television sets per capita

   PHYSTOT: Number of physicians per 1,000 people

   CTRYNAME: Name of the country

   a. What are the units of analysis in this dataset? [2 points]
   b. What is the range of each of these variables (except CTRYNAME)? [3 points]
   c. What is the mean of each variable? [3 points]
   d. Run a bivariate regression predicting infant mortality based on the number of televisions per capita.
      i. Write out the regression equation described by this analysis. [4 points]
      ii. Is the number of televisions per capita a statistically significant (at the 95% level) predictor of infant mortality? How do you know? [4 points]
      iii. What is the expected number of infant deaths per 1,000 live births in a country with 0.6 televisions per capita? [4 points]
      iv. Do you think the relationship between the number of TVs per capita and the infant mortality rate is causal? Why or why not? [4 points]
   e. Calculate the correlation between the number of TVs per capita and the number of physicians per 1,000 people.
      i. What is the correlation? [4 points]
      ii. Describe the relationship between these two variables in words. [4 points]
   f. Next, run a multivariate regression predicting infant mortality using the number of TVs per capita and the number of physicians per 1,000 people as independent variables.
      i. Write out the regression equation described by this analysis. [4 points]
      ii. Is the number of televisions per capita a statistically significant (at the 95% level) predictor of infant mortality? How do you know? What about number of physicians per 1,000 people? [6 points]
      iii. What is the expected number of infant deaths per 1,000 live births in a country with 0.4 televisions per capita and 3 physicians per 1,000 people? [4 points]

    iv. Does the coefficient on the number of televisions per capita look different in this model compared with the bivariate model (in part “d.” above)? Why might this be? Explain in words. [6 points]

g. Draw a diagram summarizing your interpretation of the causal relationships between these three variables. Describe what your diagram means in words. [6 points]

h. Think of another variable that might be confound (bias) the relationship between TVs per capita and infant mortality you estimated (in part “f.”).

    i. Describe how this variable might be measured. [4 points]

    ii. How might including this variable as an independent variable in that regression model affect the coefficient on TVs per capita? Why? [4 points]

    iii. How might including this variable affect the coefficient on physicians per 1,000 people? Why? [4 points]

    iv. Explain (in words) where this additional variable would fit in your causal diagram and why. [4 points]