

## POL 681 Problem Set 4: Categorical Variables

**Objective:** The objective of this problem set is to assess your ability to estimate and interpret regression models with dummy variables.

**Directions:** Please answer each of the questions using the information provided. SHOW ALL WORK.

The data you will use records the length of time a U.N. Peacekeeping mission has persisted. On the website, this data set is called `undata.dta`.

$Y$  denotes the dependent variable and it is measured as the *natural log* of the number of months a U.N. Peacekeeping Mission has lasted. In the Stata data set, this variable is called `logdur`.

$X_1$  denotes an independent variable having three categories. These categories indicate the type of conflict that prompted the peacekeeping mission. Category “1” denotes “civil wars”; category “2” denotes “internationalized civil wars”; and category “3” denotes interstate conflicts. In the Stata data set, this variable is called `missiontype`.

$X_2$  denotes the geographic land mass of the conflict. It is measured as the *natural log* of the actual square miles of the land mass. In the Stata data set, this variable is called `logarea`.

For this problem set, please use Stata to help you answer the following questions. Each question is worth 10 points.

1. Regress  $Y$  on  $X_1$ . What is the interpretation of the regression coefficient? What specific interpretative problems are there in estimating an OLS model using  $X_1$  as it currently is coded?
2. In order to avoid the problems delineated in question 1, recode  $X_1$  using standard dummy variable coding approaches. Reestimate the model, but now using the dummy variables. What is the precise interpretation of your results? Remember that  $Y$  is expressed in natural log units. When presenting your “substantive” results, be sure to discuss your predicted values in a metric that would be transparent to readers of your research.
3. Now, reestimate the model in 2, but this time use contrast coding. For uniformity, let “interstate conflicts” serve as the baseline category. What is the precise interpretation of your results (noting now your coding scheme has changed)? Remember that  $Y$  is expressed in natural log units. When presenting your “substantive” results, be sure to discuss your predicted values in a metric that would be transparent to readers of your research.
4. Reestimate the model in 2, but this time, include as an additional independent variable,  $X_3$ . What is the precise interpretation of  $X_3$  in this model?
5. Using graphs of predicted values from the previous model (the one estimated in 4), illustrate the “parallel slopes” assumption. Explain what this assumption implies for these data.