Data Analysis Exercises for Chapter 20: *Applied Regression Analysis, Generalized Linear Models, and Related Methods, Second Edition* (Sage, 2007)

John Fox

17 January 2008

Exercise D20.1 Using the United Nations social-indicators data (in `UnitedNations.txt`), develop a regression model for the response variable female expectation of life. Feel free to use whatever explanatory variables in the data set make sense to you, and to employ variable transformations, methods of fitting the model other than least-squares regression (e.g., robust regression), etc. Work initially with complete cases, and once you have an apparently satisfactory model, obtain estimates and standard errors of the regression coefficients by multiple imputation. Compare these results to those from the complete-case analysis. What do you conclude?

Exercise D20.2 Locate a data set for a large-scale social survey of interest to you; the survey should include a substantial amount of missing data—unfortunately, it is not hard to find surveys that fit the bill. (One source of survey data is the Inter-University Consortium for Political and Social Research data archive, at <http://www.icpsr.umich.edu/ICPSR/access/index.html>.) Specify a regression model of some sort and fit the model to the data employing a complete-case analysis. Then reestimate the parameters of the model and their standard errors by multiple imputation. Compare the multiple-imputation estimates and standard errors with those from the complete-case analysis. What do you find?

Exercise D20.3 Long (1997) reports a regression in which the response variable is the prestige of the academic departments where PhDs in biochemistry find their first jobs. Prestige is measured on a scale that runs from 1.00 to 5.00, and is unavailable for departments without graduate programs and for departments with ratings below 1.00. The explanatory variables include a dummy regressor for gender; the prestige of the department in which the individual obtained his or her PhD; the number of citations received by the individual's mentor; a dummy regressor coding whether or not the individual held a fellowship; the number of articles published by the individual; and the number of citations received by the individual. The data are in the file `Long-PhDs.txt`. Estimate the regression of prestige of first job on the other variables in three ways: (a) code all of the missing values as 1.00 and perform an OLS regression; (b) treat the missing values as truncated at 1.00 and employ Heckman's selection-regression model; and (c) treat the missing values as censored and fit the Tobit model . Finally, compare the estimates and coefficient standard errors obtained by the three approaches. Which of these approaches makes the most substantive sense?