Data Analysis Exercises for Chapter 24: *Applied Regression Analysis, Generalized Linear Models, and Related Methods, Third Edition* (Sage, 2016)

John Fox

Last modified: 2015-03-20

Exercise D24.1 In *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Cambridge University Press, 2007: Section 14.1), Gelman and Hill analyze polling data for the week prior to the 1988 U.S. presidential election. The data set in the file `Gelman.txt`, adapted from Gelman and Hill's data, contains information for respondents to CBS News polls in the 48 contiguous states; the data set includes the following variables and excludes respondents with missing data:

- `state`: The two-letter abbreviation for the state in which the respondent resided.
- `region`: Midwest, Northeast, South, or West.
- `bush`: Whether or not the respondent intended to vote for George Bush in the upcoming election (yes or no).
- `age`: The respondent's age, categorized as 18–29, 30–44, 45–64, or 65+.
- `education`: The respondent's level of education — less then high-school, high-school graduate, some college, college graduate.
- `race`: Black or white.
- `gender`: Female or male.
- `previous`: A measure of previous Republican vote in the state, expressed as a proportion, and adjusted for home-state and home-region effects in earlier presidential elections.

(a) Fit within-state binary logistic regressions of anticipated vote for Bush on age, education, race, and gender. Examine how the logistic-regression coefficients vary by state and region. Does your exploration of the data suggest a mixed-effects model for Gelman and Hill's polling data?

(b) Fit a binary logistic mixed-effects regression of anticipated vote for Bush on age, education, race, gender, and previous vote, including a random effect allowing the intercept to vary by state. You may find it easier to fit this and subsequent GLMMs if you use the logit of previous vote on the right-hand side of the model or alternatively simply subtract 0.5 from previous Republican vote. What do you find? Consider simplifying the model by removing fixed effects that are small and non-significant, and complicating it by including additional random effects.

(c) States are nested within the regional classification, and it is reasonable to expect both regional variation in voting patterns and variation by state within regions. Specify a model in which vote for Bush is regressed on the explanatory variables that you found to be important in part (b) but now include random effects for the intercept by both state and region. How, if at all, do the results vary from those in part (b)? Which model do you prefer?

Gelman and Hill's analysis of the 1988 polling data is considerably more careful and nuanced than the one pursued here, and uses Bayesian hierarchical models in addition to classical GLMMs.

Exercise D24.2 Wong, Monette, and Weiner's study of recovery from coma, in the file `Wong.txt`, collected data on verbal IQ in addition to the data on performance IQ analyzed in Section 24.2.1. Perform a similar analysis of verbal IQ: Explore the data to confirm that it is reasonable to fit an asymptotic growth model; find reasonable start values for the model; fit the nonlinear mixed-effects asymptotic growth model to the data; and interpret the results. How does the typical post-coma trajectory of recovery of verbal IQ compare to the typical trajectory of recovery of performance IQ described in the text?