Data Analysis Exercises for Chapter 15: *Applied Regression Analysis, Generalized Linear Models, and Related Methods, Third Edition* (Sage, 2016)

John Fox

Last modified: 2015-03-21

Exercise D15.1 Long (1990, 1997) investigates factors affecting the research productivity of doctoral students in biochemistry. The response variable in this investigation, art, is the number of articles published by the student during the last three years of his or her PhD programme. The explanatory variables are as follows:

fem	Gender: dummy variable — 1 if female, 0 if male
mar	Maritial status: dummy variable — 1 if married, 0 if not
kid5	Number of children five years old or younger
phd	Prestige rating of PhD department
ment	Number of articles published by mentor during last three years

Long's data (on 915 biochemists) are in the file `Long.txt`. The variable names listed above are those employed by Long.

(a) Examine the distribution of the response variable. Based on this distribution, does it appear promising to model these data by linear least-squares regression, perhaps after transforming the response? Explain your answer.

(b) Following Long, perform a Poisson regression of art on the explanatory variables. What do you conclude from the results of this regression?

(c) Perform regression diagnostics on the model fit in the previous question. If you identify any problems, try to deal with them. Are the conclusions of the research altered?

(d) Refit Long's model allowing for overdispersion (using a quasi-Poisson or negative-binomial model). Does this make a difference to the results?

Exercise D15.2 Chapter 12 describes the linear regression of wages on gender, age, and education for data drawn from the Canadian Survey of Labour and Income Dynamics (the "SLID"). The data are in the file `SLID-Ontario.txt`. In the text, the response variable is log-transformed to correct skewness and non-constant spread in the regression. Consider an alternative strategy employing a gamma generalized linear model. After fitting this model and checking its adequacy, which of the two approaches to the data do you prefer?

Exercise D15.3 Return to the logit (and probit) model that you fit in Exercise D14.1.

(a) Use the diagnostic methods for generalized linear models described in this chapter to check the adequacy of the final model that you fit to the data.

(b) If the model contains a discrete quantitative explanatory variable, test for nonlinearity by specifying a model that treats this variable as a factor (e.g., using dummy regressors), and comparing that model via a likelihood-ratio test to the model that specifies that the variable has a linear effect. (If there is more than one discrete quantitative explanatory variable, then begin with a model that treats all of them as factors, contrasting this with a sequence of models that specifies a linear effect for each such variable in turn.) Note that this is analogous to the approach for testing for nonlinearity in a linear model with discrete explanatory variables described in Section 12.4.1.

(c) Explore the use of the log-log and complementary-log-log links as alternatives to the logit link for this regression. Comparing deviances under the different links, which link appears to best represent the data?

Exercise D15.4 The file `CES11-turnout.txt` contains data from the 2011 Canadian Election Study. This is a different extract from the study used in Section 15.5 to illustrate the analysis of data from a complex survey sample. The variables in the current data set are as follows:

household	Household ID number
province	Province code (NL, PE, NS, NB, QC, ON, MB, SK, AB, BC)
population	Population of the respondent's province, number over age 17
weight	Weight sample to size of population
education	Respondent's level of education, 6 categories
gender	Female or male
age	Age in years
income	Family income, 5 categories
voted	Self-report of voter turnout, yes or no

As explained in the text, the sample was stratified by province, with disproportional sampling in the various strata, and so the first four variables pertain to the sampling design.

(a) *Model-based inference*: Ignoring the sampling design, fit a logistic regression of voter turnout on education, gender, age, and income, employing a 5-df regression spline for age. Performing appropriate statistical tests and examining the relationship of turnout to each explanatory variable, what do you conclude from the results of this logistic regression?

(b) According to Elections Canada, turnout in the 2011 Federal election was 61 percent. Compute the percentage of respondents in the 2011 Canadian Election Study who claim to have voted in the election; repeat this computation taking account of the sampling design. Do these results differ substantially from the actual turnout? If so, how might you account for the difference?

(c) *Design-based inference*: Repeat the logistic regression in part (a) but estimate the model taking the sampling design into account. Compare the results to those obtained in part (a).

(d) There is quite a bit of missing data in this data set, due, for example, to panel attrition (information on voter turnout was collected, of course, in the post-election wave of the study) and to failure to answer the income question. Using multiple imputation (described in Chapter 20) to handle the missing data, repeat the analysis in parts (a) and (c). How, if at all, do your results change?