Data Analysis Exercises for Chapter 14: *Applied Regression Analysis, Generalized Linear Models, and Related Methods, Third Edition* (Sage, 2016)

John Fox

Last modified: 2015-01-29

Exercise D14.1 Each of the following data sets contains a dichotomous response variable, along with other variables that can be treated as explanatory variables:

Data Set	*Response Variable*
`Chile.txt`	voting intention (yes vs. no)
`Greene.txt`	leave to appeal granted (yes vs. no)
`Hamilton.txt`	school-closing opinion (close vs. open)
`Mroz.txt`	labour-force participation (yes vs. no)
`HS-Powers.txt`	high-school graduation (yes vs. no)
`Titanic.txt`	survived (yes vs. no)
`CanadianWomen.txt`	labour-force participation (yes vs. no)

Alternatively, select an appropriate data set of interest to you.

(a) Formulating a model that makes substantive sense in the context of the data set — for example, constructing dummy regressors to represent factors and including interaction regressors where these are appropriate — fit a linear logistic regression of the response variable on the explanatory variables, reporting the estimated regression coefficients and their asymptotic standard errors.

(b) Construct an analysis-of-deviance table for the model fit in part (a).

(c) Fit a final model to the data that includes the statistically significant effects. Construct an effect display for each high-order term in the model. If the model is additive, (i) suggest *two* interpretations of each estimated coefficient; and (ii) construct likelihood-ratio-based 95-percent confidence intervals for the regression coefficients, comparing these with confidence intervals based on the Wald statistic.

(d) Fit a probit model to the data, comparing the results to those obtained with the logit model.

Exercise D14.2 Each of the following data sets contains a polytomous response variable, along with other variables that can be treated as explanatory variables:

Data Set	*Response Variable*
`Chile.txt`	voting intention (abstain, undecided, yes, no)
`GSS-Long.txt`	occupation (menial, blue-collar, craft, white-collar, professional)
`Women-Powers.txt`	woman's place is in the home (SD, D, A, SA)
`Moms-Long.txt`	attitude toward working mothers (SD, D, A, SA)
`CanadianWomen.txt`	labour-force participation (not working, part-time, full-time)

Family Income	*Perceived Need*	
	No	Yes
\$0	8	5
\$1–\$1999	17	16
\$2000–\$3999	88	76
\$4000–\$5999	125	108
\$6000–\$7999	134	75
\$8000–\$9999	130	94
\$10,000–\$11,999	168	79
\$12,000–\$13,999	178	65
\$14,000–\$15,999	240	60
\$16,000–\$17,999	141	45
\$18,000–\$19,999	160	45
\$20,000–\$24,999	299	71
\$25,000–\$29,999	199	29
\$30,000–\$39,999	162	23
\$40,000–\$49,999	61	7
\$50,000–\$74,999	36	3
\$75,000–\$99,999	4	0
\$100,000 or more	7	1

Table 1: *Perceived need by income.* Source of Data: *These data were collected as part of the Social Change in Canada Project, directed by T. Atkinson, M. Ornstein, and H. Stevenson of York University. The research was supported by SSHRCC grant S75-0332. The data were made available by the Institute for Social Research of York University. Neither the principal investigators nor the disseminating archive are responsible for the interpretations presented here.*

Alternatively, select a suitable data set of interest to you. Proceed as in Exercise D14.1, but use, as appropriate, one or more of the following: a multinomial logit model; a proportional-odds logit model; logit models fit to a set of nested dichotomies; or similar probit models. If you fit the proportional-odds model, test the assumption of parallel regressions. If you fit more than one kind of model, which model do you prefer? Why?

Exercise D14.3 The data shown in Table 1 are drawn from a 1977 social survey conducted by the Institute for Social Research at York University in Toronto. Respondents to the survey are classified by family income (represented by 18 categories) and by their responses to the question, "During the past year, have there been any major things you or your family *really* needed to buy but have not been able to afford?" I will refer to this second variable as "perceived need."

(a) Calculate the empirical logits, $\log_e(\text{Yes}/\text{No})$, for perceived need within income categories, and plot these logits against family income. Because there is a 0 frequency, you should add 0.5 to each frequency prior to calculating the logits. Does the relationship between perceived need and income appear to be roughly linear on the logit scale? How would you characterize the relationship between perceived need and income? In drawing the graph, use the category midpoints, expressed in thousands of dollars, as income scores, employing an arbitrary figure (say, 125) for the final, open, category.

(b) Using the binomial logit model, fit two models to the data: (i) assuming a linear relationship between the logit of perceived need and income; and (ii) coding dummy regressors for the income categories. Show the fitted line from (i) on the graph drawn in part (a). Then perform a likelihood-ratio test for nonlinearity on the logit scale.

(1) Density	*(2) Sex of Subject*	*(3) Sex of Intruder*	*(4) Response*	
			Yes	No
Low	Male	Male	18	1
		Female	15	8
	Female	Male	17	5
		Female	12	7
High	Male	Male	13	6
		Female	16	4
	Female	Male	10	9
		Female	14	6

Table 2: *Reactions to invasions of personal space.* Source of data: *Harris, Luginbuhl, and Fishbein (1978: Table 1).*

Exercise 14.4 Harris, Luginbuhl, and Fishbein (1978) conducted a social-psychological experiment that examined reactions to invasions of personal space. The research took place in a field setting provided by a public escalator. The primary results of the study were presented in the form of a contingency table (Table 2). Three of the variables in the table were design or explanatory variables: (1) density of people on the escalator, rated as either high or low; (2) the sex of the subject; and (3) the sex of the intruder. The fourth variable was a dichotomous response variable: whether or not the subject reacted in some manner to the intrusion.

The authors analyzed the data by separately examining the two-way (partial) tables relating density to response within combinations of categories of the other three variables. Because there is a statistically significant relationship between density and response in only one of the four partial tables, the authors concluded that "males in the present study were more likely to react to a personal space invasion under low-density conditions than high-density conditions, but only when the intruder was another male. Density had no effect on responses by female subjects" (Harris, Luginbuhl, and Fishbein, 1978: 352-353). The implication is that there is a three-way interaction among the explanatory variables in determining response.

(a) Calculate the response-variable odds within combinations of explanatory-variable categories. Compute and graph the log-odds (empirical logits), commenting on the results.

(b) Construct an analysis-of-deviance table, testing the various interactions and main effects of the explanatory variables on response. Do these tests square with the descriptive findings in part (a)?

(c) On the basis of the tests in part (b), fit a final logit model that incorporates only those effects shown to be important (and, of course, effects marginal to them). Using the parameter estimates for the model, calculate and graph the fitted logits.

(d) Test for independence between density and response separately in each of the four partial tables. (You may either fit a logit model to each table, or perform a traditional Pearson chi-square test of independence.) Do you obtain the results reported by Harris, Luginbuhl, and Fishbein.

(e) Do the results of your logit analysis support the authors' conclusions [replicated in part (d)]? Which analysis to you prefer? Why? (Cf., Fox, 1979.)

(1) Race	(2) County Percent Nonwhite	(3) State Suffrage Law	(4) Frequency of Voting		
			All or Most Elections	Some Elections	Never
White	< 30	Restrictive	108	58	64
		Moderate	190	74	108
	> 30	Restrictive	46	18	21
		Moderate	41	10	16
Black	< 30	Restrictive	3	5	5
		Moderate	15	8	26
	> 30	Restrictive	0	6	76
		Moderate	3	10	27

Table 3: *Electoral participation in the American South.* Source of data: *Campbell et al., (1960: Table 11-5).*

Exercise D14.5 The data in Table 3, drawn from *The American Voter*, are for a sample of respondents residing in the southern United States. All of the southern states had more or less restrictive election laws, aimed primarily at limiting the political participation of African-Americans. The table relates respondents' frequency of voting to race, the racial composition of the respondent's county of residence, and the restrictiveness of the state suffrage law.

Analyze the data in this table using:

(a) the polytomous (multinomial) logit model;

(b) dichotomous (binomial) logit models for the nested dichotomies {Never, (Some, All or Most)} and {Some, All or Most};

(c) the proportional-odds model.