Data Analysis Exercises for Chapter 3: *Applied Regression Analysis, Generalized Linear Models, and Related Methods, Third Edition* (Sage, 2016)

John Fox

Last modified: 2015-01-29

Exercise D3.1 Using the methods described in Section 3.1, examine the distributions of the quantitative variables in one of the following data sets: `Angell.txt`, `Anscombe.txt`, `Chirot.txt`, `Duncan.txt`, `Ericksen.txt`, `Freedman.txt`, `Leinhardt.txt`, `States.txt`, and `UnitedNations.txt`. Characterize the distribution of each variable in terms of symmetry or skewness; non-normality or apparent normality; number of modes; and the presence or absence of unusual values.

Exercise D3.2 Recall from Chapter 2 Davis's data on measured and reported weight for women engaged in regular exercise, plotted (for example) in Figure 2.11. Both measured and reported weight were recorded to the nearest kilogram. Correct the outlying observation, and investigate whether jittering measured and reported weight clarifies the relationship between the two variables. It might help to show the line

$$\widehat{\text{Measured weight}} = \text{Reported weight}$$

on the scatterplot.

Exercise D3.3 Each of the following data sets contains at least one quantitative response variable. Select one of these data sets, or another suitable data set of interest to you. Using the methods described in Section 3.2, examine and characterize the relationship between the response variable and each quantitative or categorical explanatory variable:

Data Set	*Suggested Response Variable*
`Angell.txt`	moral integration
`Anscombe.txt`	education expenditures
`Chirot.txt`	intensity of the rebellion
`Ericksen.txt`	undercount
`Freedman.txt`	crime
`Leinhardt.txt`	infant mortality
`States.txt`	SAT verbal or math score
`UnitedNations.txt`	total fertility rate, or expectation of life for males or females

Exercise D3.4 The data given in Table 1 (and in `Burt.txt`), on the IQs of 27 pairs of identical twins reared apart, were reported by Sir Cyril Burt (1966). (These "data" were notoriously manufactured.) One twin in each pair was presumably raised by his or her biological parents; the other twin was raised in a foster home. In each case, Burt recorded (i.e., made up) the "social class"' to which the twins' biological parents belonged. Use a scatterplot coded by

Pair	*IQ of Twin Raised by Biological Parents*	*IQ of Twin Raised by Foster Parents*	*Social Class*
1	82	82	high
2	80	90	high
3	88	91	high
4	108	115	high
5	116	115	high
6	117	129	high
7	132	131	high
8	71	78	medium
9	75	79	medium
10	93	82	medium
11	95	97	medium
12	88	100	medium
13	111	107	medium
14	63	68	low
15	77	73	low
16	86	81	low
17	83	85	low
18	93	87	low
19	97	87	low
20	87	93	low
21	94	94	low
22	96	95	low
23	112	97	low
24	113	97	low
25	106	103	low
26	107	106	low
27	98	111	low

Table 1: *Fraudulent data on IQs of identical twins reared apart, from Burt (1966).*

social class to examine the relationship between the two twins' IQs, treating the IQ of the twin raised by biological parents as the response variable. Can you tell from the plot that the data are fraudulent? (*Hint*: Compute the linear least-squares regression separately for each class and place the regression lines on the scatterplot.)

Exercise D3.5 Using the data from Exercise D3.3, construct a scatterplot matrix for the response variable and the quantitative explanatory variables. If there are more than two explanatory variable, pick two, and, along with the response variable, construct a dynamic 3D scatterplot. Characterize the pairwise relationships among the variables and the relationship of the response to the explanatory variables.