Data Analysis Exercises for Chapter 5: *Applied Regression Analysis, Generalized Linear Models, and Related Methods, Third Edition* (Sage, 2016)

John Fox

Last modified: 2015-01-29

Exercise D5.1 # The following five observations were selected from among the 20 households in Sahlins's dataset on Mazulu village:

Household	Consumers/ Gardener X_i	Acres/ Gardener Y_i
1	1.00	1.71
5	1.20	2.21
10	1.46	2.09
15	1.65	2.41
20	2.30	2.36

(The full data set is in the file `Sahlins.txt`.)

(a) Construct a scatterplot for Y and X.

(b) Find A and B for the least-squares regression of Y on X, and draw the least-squares line on the scatterplot. Interpret A and B.

(c) Calculate the standard error of the regression, S_E, and the correlation coefficient, r. Interpret these statistics.

Exercise D5.2 Analyze Sahlins's data, given in `sahlins.txt`, by regressing Acres/Gardener on Consumers/Gardener. In a society characterized by 'primitive communism,' the social product of the village would be redistributed according to need, while each household would work in proportion to its capacity, implying a regression slope of zero. In contrast, in a society in which redistribution is purely through the market, each household should have to work in proportion to its consumption needs, suggesting a positive regression slope and an intercept of zero. Interpret the results of the regression in light of these observations. Examine and interpret the values of A, B, S_E, and r (or r^2). Do the results change if the fourth household is deleted? Plot the regression lines calculated with and without the fourth household on a scatterplot of the data. Does either regression do a good job of summarizing the relationship between Acres/Gardener and Consumers/Gardener? (Cf., Exercise D2.1.)

Exercise D5.3 The data sets for the book include many that are suitable for examining the regression of one quantitative variable on another. Pick one of these data sets, or any data set from another source of interest to you, and select one or more quantitative variables to treat

as explanatory variables and another as the response. Some suggestions:

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

Draw a separate scatterplot showing the relationship of the response to each explanatory variable, and then, for each explanatory variable, compute the simple linear regression of the response on the explanatory variable, finding A, B, S_E, and r; interpret each of these quantities. Draw the least-squares line on the scatterplot. Is the least-squares line a reasonable summary of the relationship between the two variables? (Cf., Exercise D2.3.)

Exercise D5.4 # The following nine Southeastern cities are selected from Angell's data on U. S. cities (in `Angell.txt`).

City	Moral Integration Y_i	Heterogeneity X_{i1}	Mobility X_{i2}
Richmond	10.4	65.3	24.9
Chattanooga	9.3	57.7	27.2
Nashville	8.6	57.4	25.4
Birmingham	8.2	83.1	25.9
Louisville	7.7	31.5	19.4
Jacksonville	6.0	73.7	27.7
Memphis	5.4	84.5	26.7
Miami	5.1	50.2	41.8
Atlanta	4.2	70.6	32.6

Employing the data for these nine cities:

(a) Find A, B_1, and B_2 for the linear least-squares regression of Y on X_1 and X_2.

(b) Calculate the standard error of the regression, S_E, and the multiple-correlation coefficient, R.

Exercise D5.5 The data sets for the book include many that are suitable for examining the regression of one quantitative variable on two or more others, including `Angell.txt`, `Anscombe.txt`, `Chirot.txt`, `Duncan.txt`, `Ericksen.txt`, `Freedman.txt`, `States.txt`, and `UnitedNations.txt`. Pick one of these data sets, or any data set from another source of interest to you, and one quantitative variable to treat as the response and two or more quantitative variables to treat as explanatory variables. Compute the least squares regression of the response on the explanatory variables, interpreting the values that you obtain for the regression intercept A and slopes B_j, along with the standard error of the regression S_E, and the multiple-correlation coefficient R. Compare the results of the multiple regression with those of the simple regressions that you performed in Exercise D5.3.

Exercise D5.6. Compute standardized partial regression coefficients for the multiple regression that you performed in Exercise D5.5, and explain how these coefficients are to be interpreted.