Data Analysis Exercises for Chapter 2: *Applied Regression Analysis, Generalized Linear Models, and Related Methods, Third Edition* (Sage, 2016)

John Fox

Last modified: 2015-01-29

Exercise D2.1 #[1]The data in Table 1 and `Sahlins.txt` were compiled by Sahlins (1972) from information presented in Scudder's (1962) report on the Gwenba valley of Central Africa. The data describe agricultural production in Mazulu village. The explanatory variable (Consumers/Gardener) is the ratio of consumers to productive individuals in each of 20 households, making suitable adjustments for the consumption requirements of different household members. The response variable (Acres/Gardener) is a measure of domestic-labor intensity, based on the amount of land cultivated by each household. Think of Consumers/Gardener as representing the relative consumption needs of the household, and Acres/Gardener as representing how hard each productive individual in the household works. Sahlins was interested in production, consumption, and redistribution of the social product in "primitive" communities.

(a) Draw a scatterplot of Acres/Gardener (Y) versus Consumers/Gardener (X). What relationship, if any, do you discern in this plot—does the relationship appear to be positive or negative (or neither), linear or nonlinear, strong or weak? Is there anything else noteworthy about the data—for example, do any households appear to be unusual?

(b) Notice that the households are ordered by the values of Consumers/Gardener (X). Divide the 20 households into three groups, placing the first seven households in the first group, the next six in the second group, and the last seven in the third group. Calculate the mean Y and mean X in each of the three groups. Transfer these means to the scatterplot and connect them with a simple nonparametric regression line. Does the regression line help you to interpret the relationship between the variables?

(c) Two of the households, one in the first group, one in the second (which ones?), stand out from the others. Recalculate the means in groups 1 and 2 omitting the outlying observations, and redraw the nonparametric regression line. How, if at all, do the new means differ from the original values?

(d) Use a computer program to fit a nonparametric regression line to Sahlins's data using locally weighted regression (lowess). How would you summarize the relationship between Acres/Gardener and Consumers/Gardener?

Exercise D2.2 #The data in Table 2 and `Robey.txt` are drawn from a report by Robey, Shea, Rutstein, and Morris (1992) on the fertility decline in developing countries. The data show the total fertility rate (Y) and the percentage of married women aged 15 to 44 who use contraceptives (X) in each of 50 developing countries (along with the region of the world in which each country is located). The total fertility rate is the expected number of births to a

[1]Exercises intended for "hand" computation are marked with a pound sign (#).

Table 1: Data on Agricultural Production in Mazulu Village. *Source of Data:* Sahlins (1972, Table 3.1).

Household	*Consumers/ Gardener* X_i	*Acres/ Gardener* Y_i
01	1.00	1.71
02	1.08	1.52
03	1.15	1.29
04	1.15	3.09
05	1.20	2.21
06	1.30	2.26
07	1.37	2.40
08	1.37	2.10
09	1.43	1.96
10	1.46	2.09
11	1.52	2.02
12	1.57	1.31
13	1.65	2.17
14	1.65	2.28
15	1.65	2.41
16	1.66	2.23
17	1.87	3.04
18	2.03	2.06
19	2.05	2.73
20	2.30	2.36

Table 2: Fertility (Total Fertility Rate, TFR) and Contraception (Percentage of Women of Reproductive Age Who Practice Contraception) for 50 Developing Nations Around 1990. *Source of Data:* Robey, et al. (1992).

Nation	*TFR*	*Contraception*	*Nation*	*TFR*	*Contraception*
Africa			Sri Lanka	2.7	62
Botswana	4.8	35	Thailand	2.3	68
Burundi	6.5	9	Vietnam	3.9	53
Cameroon	5.9	16	*Latin America and Caribbean*		
Ghana	6.1	13	Belize	4.5	47
Kenya	6.5	27	Bolivia	4.9	32
Liberia	6.4	6	Brazil	3.6	66
Mali	6.8	5	Colombia	2.8	66
Mauritius	2.2	75	Costa Rica	3.6	70
Niger	7.3	4	Dominican Republic	3.3	56
Nigeria	5.7	6	Ecuador	3.8	53
Senegal	6.4	12	El Salvador	4.6	47
Sudan	4.8	9	Guatemala	5.6	23
Swaziland	5.0	21	Haiti	6.0	10
Tanzania	6.1	10	Jamaica	2.9	55
Togo	6.1	12	Mexico	4.0	55
Uganda	7.2	5	Panama	4.0	58
Zambia	6.3	15	Paraguay	4.6	48
Zimbabwe	5.3	45	Peru	3.5	59
Asia and Pacific			Trinidad-Tobago	3.1	54
Bangladesh	5.5	40	*Near East and North Africa*		
China	2.5	72	Egypt	4.6	40
India	4.3	45	Jordan	5.5	35
Indonesia	3.0	50	Morocco	4.0	42
Republic of Korea	1.7	77	Tunisia	4.3	51
Pakistan	5.2	12	Turkey	3.4	60
Philippines	4.3	34	Yemen	7.0	7

woman who survives through her child-bearing years under current age-specific fertility rates. Repeat Exercise D2.1 for this dataset, using three groups of 17, 16, and 17 observations.

Exercise D2.3 The data sets for the book include many that are suitable for examining the regression of one quantitative variable on another, including `Angell.txt`, `Anscombe.txt`, `Burt.txt`, `Chirot.txt`, `Duncan.txt`, `Ericksen.txt`, `Freedman.txt`, `Leinhardt.txt`, `States.txt`, and `UnitedNations.txt`. Pick one of these data sets, or any data set from another source of interest to you, and select one quantitative variable to treat as the explanatory variable and another as the response. Use a scatterplot with a nonparametric regression smoother such as lowess to examine the relationship between the two variables. What do you conclude?