

### Multiple Regression Exercise

1. Thanksgiving is this week, and many of us will overeat. But some of us are more likely to gain weight than are others. Perhaps fidgeting and other “nonexercise activity” (NEA) explain why—the body might spontaneously increase NEA the more we eat, and this effect may be bigger for some people than for others. A group of researchers<sup>1</sup> deliberately overfed 16 healthy young adults for eight weeks. They measured fat gain (in kilograms) and, as an explanatory variable, increase in energy use (in calories) from activity other than deliberate exercise—fidgeting, daily living, and the like. The data are in *nea.dta*.

- a. Analyze the relationship between fat gain and energy use and report each step of your analysis. You’re on your own now—you decide what the appropriate steps are.
- b. Could the relationship between fat gain and energy use be affected by a lurking variable? The original researchers considered the possibility that, when we overeat, our bodies might also spontaneously increase our basal metabolic rate (BMR), which measures energy use while resting. If both energy uses increased, regressing fat gain on NEA alone would be misleading.
  - i. Make a scatterplot matrix of NEA, BMR, and fat gain. Describe the relationships among the three variables. Summarize with correlations, where appropriate.
  - ii. Analyze the association between NEA, BMR, and fat gain with multiple regression. What is the regression equation? Interpret  $R^2$  and the regression coefficients. How, if at all, does adding BMR alter your analysis of fat gain?
  - iii. How well do the data satisfy the assumptions of least-squares regression? Produce and interpret diagnostic plots of residuals-versus-fitted values. Also inspect added-variable plots and report your interpretation.
  - iv. How much fat would you predict someone to gain if neither NEA nor BMR change when a person overeats by the amount the researchers set for participants in the study?
  - v. Predict the fat gain for someone whose NEA increases by 1500 calories when she overeats. Is this prediction trustworthy? Why or why not?

2. In 2001, I conducted a small survey ( $N = 100$ ) in a town along the southeastern coast of Puerto Rico. One question the data permit us to ask is whether individual attributes predict whether people report that they experience discrimination in everyday settings. Let’s look at that question in terms of three attributes: age, sex, and socioeconomic status (*gravlee.dta*).

- a. Examine the bivariate associations among all variables (hint: you may need more than one type of graph). Describe the overall patterns and any striking deviations from the patterns.

---

<sup>1</sup> Levine, J. A., N. L. Eberhardt, and M. D. Jensen. 1999. Role of Nonexercise Activity Thermogenesis in Resistance to Fat Gain in Humans. *Science* 283:212-214.

- b. Use multiple regression to estimate how discrimination is associated with age, sex, and socioeconomic status. Interpret the regression results.
- c. After fitting the regression model, construct a plot of the residuals versus fitted values and interpret the results. Is there evidence of problems with the model?
- d. Construct added-variable plots to examine the effect of potentially influential observations on individual regression coefficients.
- e. Use the postestimation command `predict` to save the residuals as a new variable, *e*. The syntax is `predict e, residuals`.
- f. Find the observation with the largest residual. A convenient way to do that is with the user-written command `hilo` (type `findit hilo` and download the package available from UCLA).
- g. Recalculate the regression, omitting the respondent with the largest residual. What effect does omitting this one case have on the model?
- h. Using the full dataset, construct a histogram of chronic discrimination. If we had started with this step, would we have done the regression with this variable in the first place?

3. James Carey studied the relationship between multiple indicators of health status and the social environment in the Nuñoa District, Peru.<sup>2</sup> The following table shows selected results from Carey's analysis.

**TABLE 9**  
*Multiple regression results for the second seasonal illness symptomatology index for Sincata households.*

Variable	Regression Coefficients	Standardized Regression Coefficients	Statistical Significance ( <i>p</i> )
1. Mean age of household members	0.41	0.44	0.01
2. Support network specialization index	-26.35	-0.44	0.01
3. Single-parent household indicator	9.63	0.32	0.03
4. Intercept	32.35		0.03

$R = 0.84; R^2 = 0.71; N = 21$  households;  $p < .0002$

The unit of analysis is the household, and the dependent variable is an index of the number of illness symptoms experienced by all members of the household. Imagine that you are Carey and are preparing these results for publication. Write a paragraph for your that summarizes the results of Table 9 in a format appropriate for publication.

<sup>2</sup> Carey, J. W. 1990. Social system effects on local-level morbidity and adaptation in the rural Peruvian Andes. *Medical Anthropology Quarterly* 4:266-295.