

## Review of Data Analysis with R: Exercises

1. If you haven't already done so, install R on your computer and verify that it works. Install the **car** package.
2. One of the data sets in **car** the package, called `States`, contains education and other data for the 50 U.S. states and Washington DC. Find out what's in the data set by looking at its help page (`?States`), and then perform a linear least-squares regression of the average SAT math score of graduating high-school students on the average teachers' salary in the states. Perform a second regression of SAT math score on both teachers' salary and percentage of students taking the SAT exam. Compare the coefficients for teachers' salary in the two regressions. How do you account for the difference? Make some graphs of the data and possibly modify your data analysis in light of what you find.
3. The data given in the data frame `Burt` in the **car** package, on the IQs of 27 pairs of identical twins reared apart, were reported by Sir Cyril Burt (1966). (These "data" are wholly fraudulent.) One twin in each (imaginary) pair was 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.
  - a. Explore the data graphically by plotting `IQbio` (as the response variable) against `IQfoster`, using a different symbol and plotting a separate linear regression line for each social class.
  - b. Then regress the IQ of the twins reared by their biological parents (`IQbio`) on the IQ of the twins reared by foster parents (`IQfoster`), dummy variables to represent the three social classes (`class`), and regressors for the interaction between foster-twin IQ and social class. [*Suggestion*: You may want to re-order the categories of the factor `class` so that they are in their natural order rather than in the (default) alphabetic order.]
  - c. Test the interaction between foster-twin IQ and social class. If the interaction proves to be non-significant, test the partial effects of foster-twin IQ and social class on biological-twin IQ. Compute the appropriate incremental F-tests using the `Anova` function in the **car** package.
  - d. How can you tell with near certainty based on your statistical analysis alone that Burt's data were 'cooked'?

4. Employing a sample of 1643 men between the ages of 20 and 24 from the U.S. National Longitudinal Survey of Youth, Powers and Xie (2000) investigate the relationship between high-school graduation and parents' education, race, family income, number of siblings, family structure, and a test of academic ability. The data set, in the file `Powers.txt` on the course web site, contains the following variables:

<code>hsgrad</code>	Whether the respondent was graduated from high school by 1985 (Yes or No)
<code>nonwhite</code>	Whether the respondent is black or Hispanic (Yes or No)
<code>mhs</code>	Whether the respondent's mother is a high-school graduate (Yes or No)
<code>fhs</code>	Whether the respondent's father is a high-school graduate (Yes or No)
<code>income</code>	Family income in 1979 (in \$1000s) adjusted for family size
<code>asvab</code>	Standardized score on the Armed Services Vocational Aptitude Battery <sup>1</sup>
<code>nsibs</code>	Number of siblings
<code>intact</code>	Whether the respondent lived with both biological parents at age 14 (Yes or No)

- a. Following Powers and Xie perform a logistic regression of `hsgrad` on the other variables in the data set.
  - b. The logistic regression in (a.) assumes that the partial relationship between the log-odds of high-school graduation and number of siblings is linear. Test for nonlinearity by fitting a model that treats `nsibs` as a factor, performing an appropriate likelihood-ratio test. In the course of working this problem, you should discover two errors in the data. Deal with the errors in a reasonable manner. Does the result of the test change?
5. Long (1990, 1997) investigates factors affecting the research productivity of doctoral students in biochemistry. The response variable in this investigation, `art`, is the number of articles published by the student during the last three years of his or her PhD programme. The explanatory variables are as follows:

<code>fem</code>	dummy variable: 1 if female, 0 if male
<code>mar</code>	dummy variable: 1 if married, 0 if not
<code>kid5</code>	number of children five years old or younger
<code>phd</code>	prestige of PhD department
<code>ment</code>	number of articles published by mentor during last three years

Long's data (on 915 biochemists) are in the file `Long.txt`, available on the course web site. The variable names listed above are those employed by Long, and appear in the first row of the data file.

---

<sup>1</sup> Apparently scores on this test were standardized to a mean of 0 and standard deviation of 1 for the NLSY sample as a whole; as you can verify, the mean and standard deviation in this subsample differ somewhat from 0 and 1, respectively

- a. Examine the distribution of the response variable, `art`. Based on this distribution, does it appear promising to model these data by linear least-squares regression, perhaps after transforming the response?
- b. Following Long, perform a Poisson regression of `art` on the explanatory variables.
- c. Refit Long's model allowing for overdispersion (i.e., using the `quasipoisson` family in R). Does this make a difference to the results?