Introduction to the R Statistical Computing Environment
Statistical Models in R: Exercises

John Fox
(McMaster University)
ICPSR Summer Program

2010

1. 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 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. See `?Burt` for more information.

1. 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. *Hint:* You can use the **car** command `scatterplot(IQbio ~ IQfoster | class, data=Burt, smooth=FALSE)` to make this graph.
2. 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.
3. 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.
4. Based solely on your statistical analysis of the data, how can you tell with a high level of certainty that the data are "cooked"?

2. 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 workshop web site, contains the following variables:

`hsgrad`	high-school graduate by 1985 (`Yes` or `No`)
`nonwhite`	black or Hispanic (`Yes` or `No`)
`mhs`	mother is a high-school graduate (`Yes` or `No`)
`fhs`	father is a high-school graduate (`Yes` or `No`)
`income`	family income in 1979 ($1000s) adjusted for family size
`asvab`	score on the Armed Services Vocational Aptitude Battery
`nsibs`	number of siblings
`intact`	lived with both biological parents at age 14 (`Yes` or `No`)

Following Powers and Xie, perform a logistic regression of `hsgrad` on the other variables in the data set. This logistic regression 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?

3. 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:

`fem`	dummy variable: 1 if female, 0 if male
`mar`	dummy variable: 1 if married, 0 if not
`kid5`	number of children five years old or younger
`phd`	prestige of PhD department (score from 0.76 to 4.62)
`ment`	number of articles published by mentor in last three years

Long's data (on 915 biochemists) are in the file `Long.txt`, available on the workshop web site. The variable names listed above are those employed by Long, and appear in the first row of the data file (not, by the way, in the order given above).

1. 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?
2. Following Long, perform a Poisson regression of `art` on the explanatory variables.
3. Refit Long's model allowing for overdispersion (e.g., using the `quasipoisson` family). Does this make a difference to the results?

4. Winer's venerable 1971 text *Statistical Principles in Experimental Design, Second Edition* contains data from a "modified version" of an experiment attirbuted to Meyer and Noble (1958): Six subjects high in anxiety and six low in anxiety were randomly assigned to two conditions of muscular tension (no tension and high tension), yielding three subjects in each combination of conditions of anxiety and tension. The response variable is the number of errors on a learning task made by the subjects during four trial blocks of the experiment. Perform a repeated-measures analysis of variance of the data, which are in the file `Winer.txt` on the workshop web site, where the variables are named `anxiety`, `tension`, `errors.1`, `errors.2`, `errors.3`, and `errors.4`. Graph the mean number of errors as a function of anxiety, tension, and trial blocks.