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

John Fox
(McMaster University)
ICPSR

2018

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 lectures-series 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:

`gender`	factor: `female` or `male`
`married`	factor: `yes` or `no`
`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 lecture-series 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. Examine the adequacy of one or more of the statistical models fit in problems 1–3.

5. The file `Goldstein.txt` on the lecture-series website contains data on 728 11-year-old students in 48 inner-London primary schools. The data are analyzed by Harvey Goldstein in *Multilevel Statistical Models, Third Edition* (Arnold, 2003). The data set includes the following variables:

- `math.8`: a math-test score when the student was eight years old.
- `math.11`: a current math-test score.
- `female`: a dummy variable coded 1 for girls and 0 for boys.
- `manual`: a dummy variable coded 1 if the student's parent (presumably the main wage earner) is in a manual occupation and 0 otherwise.
- `school`: a number (ranging from 1 to 50) indicating which school the student attends. (Yes, there are only 48 schools!)

Add the following two variables to the data set:

1. the mean age-8 math score in the student's school;
2. the deviation between the student's own age-8 math score and the mean score in his or her school (i.e., compute the school-centered age-8 math score).

If you have difficulty creating these variables and adding them to the data set, you will find the necessary R code in the file `Goldstein.R` on the workshop web site.

1. Using Trellis graphics (i.e., the R `lattice` package), examine scatterplots of age-11 math score by centered age-8 math score for each school. Do these relationships seems reasonable linear? Note that some schools have very small numbers of observations and none has very many; it therefore isn't useful to plot nonparametric-regression smooths on the scatterplots. Then examine the relationship between age-11 math score and gender, and between age-11 math score and "social class." (If you have trouble formulating these graphs, the requisite R code is in `Goldstein.R`.)
2. Using the `lmList` function in the `nlme` package, regress age-11 math scores on centered age-8 scores and the dummy variables for gender and class. Look at the within-schools coefficients. Why are some missing? Then plot each set of coefficients (i.e., starting with the intercepts) against the schools' mean age-8 math scores. Do the coefficients appear to vary systematically by the school's mean age-8 scores? (Once again, you'll find R commands for these computations and graphs in `Goldstein.R`.)
3. Fit linear mixed-effects models to the Goldstein data (using `lmer` in the `lme4` package), proceeding as follows:
    - Begin with a one-way random-effects ANOVA of age-11 math scores by schools. How much of the variation in age-11 scores is between schools (i.e., what is the intra-class correlation)?
    - Fit a random-coefficients regression of age-11 math scores on the students' centered grade-8 scores, gender, and class. Initially include random effects for the intercept and all three explanatory variables. Test whether each of these random effects is needed and eliminate from the model those that are not. How, if at all, are grade-11 math scores related to the three explanatory variables? *Note*: Some of these mixed models take awhile to converge.

- Introduce the mean school age-8 math score as a level-2 explanatory variable, but only for the level-1 coefficients that were found to vary significantly among schools in part (b). Test whether the random effects that are in the model are still required now that there is a level-2 predictor in the model.
- Briefly summarize your findings.