Exercises Modelling

October 12, 2015

1 School Attainment

1.1 Data

Garner and Raudenbush (1991) and Raudenbush and Bryk (2002) studied the role of school and neighbourhood effects on educational attainment. The data set they used (neighbourhood.tab) was for young people who left school between 1984 and 1986 from one Scottish Educational authority. The same data were used by Rabe-Hesketh and Skrondal (2005).

1.2 Codebook

- neighid: respondents neighbourhood identifier
- schid: respondents schools identifier
- attain: respondents combined end of school educational attainment as measured by grades from various exams
- p7vrq: respondents verbal reasoning quotient as measured by a test at age 11-12 in primary school
- p7read: respondents reading test score as measured by a test at age 11-12 in primary school
- dadocc: respondents fathers occupation
- dadunemp: 1 if respondents father unemployed, 0 otherwise
- daded: 1 if respondents father was in full time education after age 15, 0 otherwise

- momed: 1 if respondents mother was in full time education after age 15, 0 otherwise
- male: 1 if respondent is male, 0 otherwise
- deprive: index of social deprivation for the local community in which the respondent lived
- dummy: 1 to 4; representing collections of the schools or neighbourhoods

1.3 Exercises

1. load the data `neighborhood.tab` using the `read.table()` command
   - how many different schools? how many persons per school?
   - how many different neighborhoods? how many persons per neighborhood?
   - what is the mean of `attain` per school (use `tapply()`)
2. Estimate a linear model on attainment (`attain`) without covariates (null-model), parameter estimates and standard errors; interpret the parameter estimates
3. Define a second model, allow for the school random effect (`schid`). Assuming the name of the model is `m1.2` type

   ```
   > coef(m1.2)
   ```

   What does the output mean? How do you get the corresponding parameter to the estimate from the model without random effect? Do they differ? Why or why not?
4. visualize the attainment score in different scores using an appropriate plot.
5. Add the observed student specific effects (gender, verbal reasoning quotient, reading ability). How does the magnitude of the school random effect change? (use `summary()` and look for the respective information)
6. use `neighid` instead of `schid` and repeat the analysis
7. What do the results of using either the `schid` or the `neighid` random effects tell you about what effects are needed in the modelling of attainment with this data set? (use `confint()` to have a look at the confidence interval of the between standard deviation of schools resp. neighborhoods

2 Binary Model of Trade Union Membership

2.1 Data

Vella and Verbeek (1998) analysed the male data from the Youth Sample of the US National Longitudinal Survey for the period 1980-1987. The number of young males in the sample is 545. The version of the data set (wagepan.tab) we use was obtained from Wooldridge (2002). The same data were used for modelling the binary response trade union membership by Rabe-Hesketh and Skrondal (2005, exercise 4.7).

2.2 Codebook

- nr: person identifier
- year: 1980 to 1987
- black: 1 if respondent is black,0 otherwise
- exper: labour market experience (age-6-educ)
- hisp: 1 if respondent is Hispanic, 0 otherwise
- poorhlth: 1 if respondent has a health disability, 0 otherwise
- married: 1 if respondent is married, 0 otherwise
- nrthcen: 1 if respondent lives in the Northern Central part of the US, 0 otherwise
- nrtheast: 1 if respondent lives in the North East part of the US, 0 otherwsie
- rur: 1 if respondent lives in a rural area, 0 otherwise
- south: 1 if respondent lives in the South of the US, 0 otherwise
- educ: years of schooling
- union: 1 if the respondent is a member of a trade union, 0 otherwise
- d8m: 1 if the year is 198m, 0 otherwise, m=1,. . . ,7

2.3 Exercises

1. load the data
   - how many observations?
   - how many persons? How many rows per person?
   - what is the percentage of black?
   - what the percentage of Hispanic?
   - what is the percentage of married?
2. Estimate a logit (binary) model (`glm()`) for trade union membership (union), without covariates, obtain parameter estimates and standard errors.
3. now use the `glmer()` function (which is the generalized form of the `lmer()`) and add the person identifier as random effect. How does the estimate change? Is this random effect significant? (use `confint()`)
4. add year as a covariate. interpret the coefficients.