Exercises for sabreR (Sabre in R)

Version 2 (Draft)

email: r.crouchley@lancaster.ac.uk

March 19, 2009

Abstract

When we use the term mass 20 in the test of one of the exercises, its is our shorthand for the fact that you will need to use 20 quadrature points, e.g. `first.mass=20` in the sabre(.,.,.) R command.

Contents

1 Exercise C1. Linear Model of Essay Grading

Johnson and Albert (1999) analysed data on the grading of essays by several experts. Essays were graded on a scale of 1 to 10 with 10 being excellent. In this exercise we use the subset of the data limited to the grades from graders 1 and 4 on 198 essays (`grader1.tab`). The same data were used by Rabe-Hesketh and Skrondal (2005, exercise 1.5).

1.1 Data description for `grader1.tab`

Number of observations (rows): 198
Number of level-2 cases: 198

1.2 Variables

`grade1`: grade awarded by grader 1 {1,2,...,10)
`grade4`: grade awarded by grader 4 {1,2,....,10}
`essay`: essay identifier

grade1	grade4	essay
8	10	1
7	5	2
2	1	3
5	5	4
7	7	5
10	10	6
5	7	7
2	3	8
5	5	9
7	4	10
5	4	11
7	7	12
5	9	13

The first few lines of `grader1.tab`

To use the data in sabreR we need to stack the data, with `grade1` and `grade4` as a single column grade. We have done this for you and generated an identifier to distinguish grade1 and grade4, i.e. `dg4`=1, if grade4 =1 and 0 otherwise.

1.3 Data description for `grader2.tab`

Number of observations (rows): 396
Number of level-2 cases: 198

1.4 Variables

`ij`: essay identifier (1,2,...,198)
`r`: response (1,2)
`grade`: grade awarded
`essay`: essay identifier (this is a copy of ij)
`dg1`: 1 if this is the grade from grader 1, 0 otherwise
`dg4`: 1 if this is the grade from grader 4, 0 otherwise

ij	r	grade	essay	dg1	dg4
1	1	8	1	1	0
1	2	10	1	0	1
2	1	7	2	1	0
2	2	5	2	0	1
3	1	2	3	1	0
3	2	1	3	0	1
4	1	5	4	1	0
4	2	5	4	0	1
5	1	7	5	1	0
5	2	7	5	0	1
6	1	10	6	1	0
6	2	10	6	0	1
7	1	5	7	1	0
7	2	7	7	0	1
8	1	2	8	1	0
8	2	3	8	0	1
9	1	5	9	1	0
9	2	5	9	0	1
10	1	7	10	1	0
10	2	4	10	0	1
11	1	5	11	1	0

The first few lines of `grader2.tab` (the stacked version of data)

1.5 Suggested exercise

1. Estimate the linear model using sabreR on `grade`, with just a constant and no other effects, obtain the log likelihood, parameter estimates and standard errors. These results are given as the 1st part (homogeneous model) of the output that is obtained by estimating the random effects model. This is done in Task 2.

2. Estimate the linear model, allowing for the essay random effect, use mass 20. Are the essay effects significant? What impact do they have on the model? Try using adaptive quadrature to see if fewer mass points are needed.

3. Re-estimate the linear model allowing for both the essay random effect and `dg4`, use adaptive quadrature with an increasing number of mass points until likelihood convergence occurs.

4. How do the results change as compared to a model with just a constant? Interpret your results.

1.6 References

Johnson, V. E., and Albert, J., H., (1999), Ordinal Data Modelling, Springer, New York.

Rabe-Hesketh, S., and Skrondal, A., (2005), Multilevel and Longitudinal Modelling using Stata, Stata Press, Stata Corp, College Station, Texas.

2 Exercise C2. Linear Model of Educational Attainment

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, exercise 2.2).

2.1 Data description for `neighbourhood.tab`

Number of observations (rows): 2310
Number of level-2 cases: 17 (`schid`); 524 (`neighid`)

2.2 Variables

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

neighid	schid	attain	p7vrq	p7read	dadocc	dadunemp	daded	momed	male	deprive	dummy
675	0	0.74	21.97	12.13	2.32	0	0	0	1	-0.18	1
647	0	0.26	-7.03	-12.87	16.20	0	0	1	0	0.21	1
650	0	-1.33	-11.03	-31.87	-23.45	1	0	0	1	0.53	1
650	0	0.74	3.97	3.13	2.32	0	0	0	1	0.53	1
648	0	-0.13	-2.03	0.13	-3.45	0	0	0	0	0.19	1
648	0	0.56	-5.03	-0.87	-3.45	0	0	0	0	0.19	1
665	0	-0.36	-2.03	-1.87	16.20	0	0	0	1	0.38	1
661	0	0.74	8.97	3.13	2.32	0	0	0	0	-0.40	1
675	0	-0.36	-2.03	4.13	-3.45	0	1	1	1	-0.18	1
664	0	0.91	16.97	28.13	-3.45	0	0	1	0	-0.17	1
663	0	0.16	-4.03	-8.87	-9.09	0	0	0	1	-0.22	1
661	0	1.52	17.97	25.13	2.32	0	0	0	0	-0.40	1
665	0	0.26	5.97	7.13	-11.49	1	0	0	0	0.38	1
668	0	0.03	0.97	-11.87	2.32	0	0	0	0	-0.24	1
687	0	-0.13	6.97	12.13	-11.49	0	0	0	1	-0.05	1

The first few lines of `neighbourhood.tab`

We can use both the school identifier (`schid=0,1,2,...,20`) and the neighbourhood identifier (`neighid`) as alternative level-2 random effects in this data set.

2.3 Suggested exercise

1. Estimate a linear model on attainment (`attain`) without covariates, obtain the log likelihood, parameter estimates and standard errors. These results are given as the 1st part (homogeneous model) of the output that is obtained by estimating the random effects model. This is done in Task 2.

2. Allow for the school random effect (`schid`), use adaptive quadrature with mass 4. Is this random effect significant?

3. Add the observed student specific effects, increase the number of mass points until the likelihood converges. How does the magnitude of the school random effect change?

4. Add the neigbhourhood effect (`deprive`). Check the number of mass points required. How does the magnitude of the school random effect change?

5. A data set sorted by the neighbourhood identifier (`neighid`); has been made available for you, this data set is called `neighbourhood2.tab`. Reestimate the constant only model allowing for neighbourhood random effect (`neighid`), use adaptive quadrature with mass 12. Is there a significant `neighd` random effect?

6. Add the student specific effects, how does the magnitude of the `neighid` random effect change?

7. Add observed neighbourhood effect `deprive` to the model, how does the magnitude of the `neighid` random effect change?

8. 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?

9. What do the two sets of results show/suggest?

2.4 References

Garner, C. L., and Raudenbush, S. W., (1991), Neighbourhood effects on educational attainment: A multilevel analysis of the influence of pupil ability, family, school and neighbourhood, Sociology of education, 64, 252-262.

Raudenbush, S. W., and Bryk, A. S., (2002), Hierarchical Linear Models, Sage, Cityplace Thousand Oaks, State CA.

Rabe-Hesketh, S., and Skrondal, A., (2005), Multilevel and Longitudinal Modelling using Stata, Stata Press, Stata Corp, College Station, Texas.

3 Exercise C3. Binary Response Model of Essay Grades

Johnson and Albert (1999) analysed data on the grading of the same essay by five experts. Essays were graded on a scale of 1 to 10 with 10 being excellent. In this exercise we use the subset of the data limited to the grades from graders 1 to 5 on 198 essays (`essays2.tab`). The same data were used by Rabe-Hesketh and Skrondal (2005, exercise 5.4).

3.1 Data description for `essays2.tab`

Number of observations: (rows): 990
Number of level-2 cases: 198

3.2 Variables

`essay`: essay identifier (1,2,...,198}
`grader`: grader identifier {1,2,3,4,5}
`grade`: essay grade {1,2,...,10}
`rating`: essay rate {1,2,...,10}, not used in this exercise
`constant`: 1 for all observations, not used in this exercise
`wordlength`: average word length
`sqrtwords`: square root of the number of words in the essay
`commas`: number of commas times 100 and divided by the number of words in the essay
`errors`: percentage of spelling errors in the essay
`prepos`: percentage of prepositions in the essay
`sentlength`: average length of sentences in the essay
`pass`: 1, if grade (5-10), 0 if grade (1-4)
`grader2`: 1, if grader =2, 0 otherwise
`grader3`: 1, if grader =3, 0 otherwise
`grader4`: 1, if grader =4, 0 otherwise
`grader5`: 1, if grader =5, 0 otherwise

essay	grader	grade	rating	constant	wordlength	sqrtwords	commas	errors	prepos	sentlength	pass	grader2	grader3	grader4	grader5
1	3	8	8	1	4.76	15.46	5.60	5.55	8	19.53	1	0	1	0	0
1	1	8	8	1	4.76	15.46	5.60	5.55	8	19.53	1	0	0	0	0
1	4	8	8	1	4.76	15.46	5.60	5.55	8	19.53	1	0	0	1	0
1	2	6	8	1	4.76	15.46	5.60	5.55	8	19.53	1	1	0	0	0
1	5	5	8	1	4.76	15.46	5.60	5.55	8	19.53	1	0	0	0	1
2	2	5	7	1	4.24	9.06	3.60	1.27	9.5	16.38	1	1	0	0	0
2	4	5	7	1	4.24	9.06	3.60	1.27	9.5	16.38	1	0	0	1	0
2	3	3	7	1	4.24	9.06	3.60	1.27	9.5	16.38	0	0	1	0	0
2	1	7	7	1	4.24	9.06	3.60	1.27	9.5	16.38	1	0	0	0	0
2	5	3	7	1	4.24	9.06	3.60	1.27	9.5	16.38	0	0	0	0	1
3	5	1	2	1	4.09	16.19	1.10	2.61	14	18.43	0	0	0	0	1
3	1	2	2	1	4.09	16.19	1.10	2.61	14	18.43	0	0	0	0	0
3	4	1	2	1	4.09	16.19	1.10	2.61	14	18.43	0	0	0	1	0
3	2	1	2	1	4.09	16.19	1.10	2.61	14	18.43	0	1	0	0	0
3	3	1	2	1	4.09	16.19	1.10	2.61	14	18.43	0	0	1	0	0
4	4	5	5	1	4.36	7.55	1.80	1.81	0	14.65	1	0	0	1	0
4	5	3	5	1	4.36	7.55	1.80	1.81	0	14.65	0	0	0	0	1
4	1	5	5	1	4.36	7.55	1.80	1.81	0	14.65	1	0	0	0	0

The first few lines of `essays2.tab`

3.3 Suggested exercise

1. Fit a binary probit model to the binary response `pass,` but without any random effects, obtain the log likelihood, parameter estimates and standard errors. These results are given as the 1st part (homogeneous model) of the output that is obtained by estimating the random effects model. This is done in Task 2.

2. Fit a binary probit model allowing for the `essay` random effect, is the `essay` effect significant? How many adaptive quadrature points should we use to estimate this model?

3. Add the 4 grader dummy variables to the model, what are the differences between the graders?

4. Add the 6 essay characteristics (`wordlength-sentlength)` to the previous model. Which of them are significant? How has including the essay characteristics improved the model?

5. Create interaction effects between the grader specific dummy variables and the `sqrtwords` explanatory variable and add these effects to the model. What do the results tell you?

3.4 References

Johnson, V. E., and Albert, J. H., (1999), Ordinal Data Modelling, Springer, New York.

Rabe-Hesketh, S., and Skrondal, A., (2005), Multilevel and Longitudinal Modelling using Stata, Stata Press, Stata Corp, College Station, Texas.

4 Exercise C4. Ordered Response Model of Essay Grades

Johnson and Albert (1999) analysed data on the grading of the same essay by five experts. Essays were graded on a scale of 1 to 10 with 10 being excellent. In this exercise we use the subset of the data limited to the grades from graders 1 to 5 on 198 essays (`essays_ordered.tab`). The same data were used by Rabe-Hesketh and Skrondal (2005, exercise 5.4) and in Exercise C3, where `grade` was recoded into a binary response. In this exercise we use `grade` as the ordered response `ngrade` with 4 categories.

4.1 Data description for `essays_ordered.tab`

Number of observations (rows): 990
Number of level-2 cases: 198

4.2 Variables

`essay`: essay identifier (1,2,...,198}
`grader`: grader identifier {1,2,3,4,5}
`grade`: essay grade {1,2,...,10}
`rating`: essay rate {1,2,...,10}, not used in this exercise
`constant`: 1 for all observations, not used in this exercise
`wordlength`: average word length
`sqrtwords`: square root of the number of words in the essay
`commas`: number of commas times 100 and divided by the number of words in the essay
`errors`: percentage of spelling errors in the essay
`prepos`: percentage of prepositions in the essay
`sentlength`: average length of sentences in the essay
`grader2`: 1 if grader =2, 0 otherwise
`grader3`: 1 if grader =3, 0 otherwise
`grader4`: 1 if grader =4, 0 otherwise
`grader5`: 1 if grader =5, 0 otherwise
`ngrade`: 1 if grade (1,2), 2 if grade (3,4); 3 if grade (5,6} and 4 if grade (7,8,9,10}

essay	grader	grade	rating	cons	wordlength	sqrtwords	commas	errors	prepos	sentlength	pass	grader2	grader3	grader4	grader5	ngrade
1	3	8	8	1	4.76	15.46	5.60	5.55	8.00	19.53	1	0	1	0	0	4
1	1	8	8	1	4.76	15.46	5.60	5.55	8.00	19.53	1	0	0	0	0	4
1	4	8	8	1	4.76	15.46	5.60	5.55	8.00	19.53	1	0	0	1	0	4
1	2	6	8	1	4.76	15.46	5.60	5.55	8.00	19.53	1	1	0	0	0	3
1	5	5	8	1	4.76	15.46	5.60	5.55	8.00	19.53	1	0	0	0	1	3
2	2	5	7	1	4.24	9.06	3.60	1.27	9.50	16.38	1	1	0	0	0	3
2	4	5	7	1	4.24	9.06	3.60	1.27	9.50	16.38	1	0	0	1	0	3
2	3	3	7	1	4.24	9.06	3.60	1.27	9.50	16.38	0	0	1	0	0	2
2	1	7	7	1	4.24	9.06	3.60	1.27	9.50	16.38	1	0	0	0	0	4
2	5	3	7	1	4.24	9.06	3.60	1.27	9.50	16.38	0	0	0	0	1	2
3	5	1	2	1	4.09	16.19	1.10	2.61	14.00	18.43	0	0	0	0	1	1
3	1	2	2	1	4.09	16.19	1.10	2.61	14.00	18.43	0	0	0	0	0	1
3	4	1	2	1	4.09	16.19	1.10	2.61	14.00	18.43	0	0	0	1	0	1
3	2	1	2	1	4.09	16.19	1.10	2.61	14.00	18.43	0	1	0	0	0	1
3	3	1	2	1	4.09	16.19	1.10	2.61	14.00	18.43	0	0	1	0	0	1
4	4	5	5	1	4.36	7.55	1.80	1.81	0.00	14.65	1	0	0	1	0	3
4	5	3	5	1	4.36	7.55	1.80	1.81	0.00	14.65	0	0	0	0	1	2
4	1	5	5	1	4.36	7.55	1.80	1.81	0.00	14.65	1	0	0	0	0	3
4	3	4	5	1	4.36	7.55	1.80	1.81	0.00	14.65	0	0	1	0	0	2
4	2	3	5	1	4.36	7.55	1.80	1.81	0.00	14.65	0	1	0	0	0	2

The first few lines of `essays_ordered.tab`

4.3 Suggested exercise

1. Fit an ordered probit model to `ngrade` but without any random effects, obtain the log likelihood, parameter estimates and standard errors. These results are given as the 1st part (homogeneous model) of the output that is obtained by estimating the random effects model. This is done in Task 2.

2. Fit an ordered probit model allowing for the `essay` random effect, is the `essay` effect significant? How many adaptive quadrature points should we use to estimate this model?

3. Add the dummy variables for `graders` (2,3,4,5) to the model, are there differences between the graders?

4. Add the 6 essay characteristics (`wordlength-sentlength`) to the previous model. Which of them are significant? Has including the essay characteristics improved the model?

5. Create interaction effects between the `grader` specific dummy variables and the `sqrtwords` explanatory variable and add these effects to the model. What do the results tell you?

6. Repeat exercise components 2-6 treating `grade` as an ordered probit model with all the observed categories (1,2,. . . ,8) of `grade`, grades (9,10) are not observed in this data set.

7. Are there any differences between the results obtained using the alternative ordered responses `ngrade` and `grade`? What does this tell you?

4.4 References

Johnson, V. E., and Albert, J. H., (1999), Ordinal Data Modelling, Springer, StateplaceNew York.

Rabe-Hesketh, S. and Skrondal, A., (2005), Multilevel and Longitudinal Modelling using Stata, Stata Press, Stata Corp, College Station, Texas.

5 Exercise C5. Poison Model of Headaches

McKnight and van den Eeden (1993) and Hedeker (1999) analysed some multi-period, two treatment crossover data (`headache2.tab`) to establish whether the artificial sweetener (`aspartame`) caused headaches. The trial involved randomly assigning 27 patients to different sequences of placebo and aspartame. We ignore the crossover aspect of the trial in this exercise. The same data were used by Rabe-Hesketh and Skrondal (2005, exercise 6.2).

5.1 Data description for `headache2.tab`

Number of observations (rows): 122
Number of level-2 cases: 27

5.2 Variables

`id`: subject identifier (1,2,. . . ,27)
`y`: count of number of headaches over several days
`cons`: 1 for all rows (not used in this analysis)
`aspartame`: 1 if treatment was aspartame, 0 otherwise
`days`: number of days for which the headaches were counted, which takes the values (1,2,. . . ,7)

id	y	cons	aspartame	days
2	0	1	0	7
2	5	1	1	7
2	2	1	0	7
5	3	1	0	7
5	0	1	1	7
5	2	1	0	7
5	0	1	1	7
5	0	1	0	7
13	7	1	0	7
13	7	1	1	7
13	7	1	0	7
13	6	1	1	7
13	7	1	0	7
16	1	1	0	7
16	3	1	1	7
16	1	1	0	7
19	0	1	0	7

The first few lines of `headache2.tab`

5.3 Suggested exercise

1. Use the offset `lt`=log(`days`) in the following Tasks.

2. Fit a Poisson model to `y` (number of headaches) with a log link without any `id` random effects, obtain the log likelihood, parameter estimates and standard errors. These results are given as the 1st part (homogeneous model) of the output that is obtained by estimating the random effects model. This is done in Task 3.

3. Fit a Poisson model to `y` allowing for the `id` random effect. Is the id random effect significant? How many adaptive quadrature points should we use to estimate this model?

4. Add the treatment indicator `aspartame` to the previous model, is there a significant treatment effect?

The responses are actually in temporal order, but we do not use that feature of the data here. Hedeker found no evidence of a sequence effect.

5.4 References

Hedeker, D., (1999), MIXNO: A computer program for mixed effects logistic regression, Journal of Statistical Software, 4, 1-92.

McKnight, B., and van den Eeden, S. K., (1993) A conditional analysis for two treatment multiple-period crossover design with binomial or Poisson outcomes and subjects who drop out, Statistics in Medicine, 12, 825-834.

Rabe-Hesketh, S., and Skrondal, A., (2005), Multilevel and Longitudinal Modelling using Stata, Stata Press, Stata Corp, College Station, Texas.

6 Exercise L1. Linear Model of Psychological Distress

Dunn (1992) reported data for the 12-item version of Goldberg's (1972) General Health Questionnaire for psychological distress. The questionnaire was completed by 12 students on 2 dates, 3 days apart. The data `ghq2.tab` are repeated in the table below, the same data were used by Rabe-Hesketh and Skrondal (2005, exercise 1.2).

6.1 Data description for `ghq2.tab`

Number of observations (rows): 24
Number of level-2 cases: 12

6.2 Variables

`ij`: student identifier
`r`: response occasion 1, 2
`student`: student identifier {1,2,...,12}
`ghq`: psychological distress score at occasion
`dg1`: 1, if the response occasion is 1, 0 otherwise
`dg2`: 1, if the response occasion is 2, 0 otherwise

ij	r	student	ghq	dg1	dg2
1	1	1	12	1	0
1	2	1	12	0	1
2	1	2	8	1	0
2	2	2	7	0	1
3	1	3	22	1	0
3	2	3	24	0	1
4	1	4	10	1	0
4	2	4	14	0	1
5	1	5	10	1	0
5	2	5	8	0	1
6	1	6	6	1	0
6	2	6	4	0	1
7	1	7	8	1	0
7	2	7	5	0	1
8	1	8	4	1	0
8	2	8	6	0	1
9	1	9	14	1	0
9	2	9	14	0	1

First few lines of `ghq2.tab`

6.3 Suggested exercise

1. Estimate the linear model in sabre on `ghq`, with just a constant, and no random effects, obtain the log likelihood, parameter estimates and standard errors. These results are given as the 1st part (homogeneous model) of the output that is obtained by estimating the random effects model. This is done in Task 2.

2. Estimate the linear model, allowing for the student random effect, use adaptive quadrature with mass 12. Are the `student` random effects significant? What does the significance mean? What impact do the `student` random effects have on the model?

3. Re-estimate the linear model allowing for both `student` random effects and `dg2`. How do the results change (compared to part 2)?

6.4 References

Dunn, G., (1992), Design and analysis of reliability studies, Statistical Methods in Medical Research, 1, 123-157.

Rabe-Hesketh, S., and Skrondal, A., (2005), Multilevel and Longitudinal Modelling using Stata, Stata Press, Stata Corp, College Station, Texas.

7 Exercise L2. Linear Model of log Wages

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). Here we study the determinants of wages. The same data were used by Rabe-Hesketh and Skrondal (2005, exercise 2.7).

7.1 Data description for `wagepan.tab`

Number of observations (rows): 4360
Number of level-2 cases: 545

7.2 Variables

`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
`lwage`: log of hourly wage in US dollars

nr	year	agric	black	bus	construc	ent	exper	fin	hisp
13	1980	0	0	1	0	0	1	0	0
13	1981	0	0	0	0	0	2	0	0
13	1982	0	0	1	0	0	3	0	0
13	1983	0	0	1	0	0	4	0	0
13	1984	0	0	0	0	0	5	0	0
13	1985	0	0	1	0	0	6	0	0
13	1986	0	0	1	0	0	7	0	0
13	1987	0	0	1	0	0	8	0	0
17	1980	0	0	0	0	0	4	0	0
17	1981	0	0	0	0	0	5	0	0
17	1982	0	0	0	0	0	6	0	0
17	1983	0	0	0	0	0	7	0	0
17	1984	0	0	0	0	0	8	0	0
17	1985	0	0	0	1	0	9	0	0
17	1986	0	0	0	1	0	10	0	0
17	1987	0	0	0	1	0	11	0	0
18	1980	0	0	0	0	0	4	0	0
18	1981	0	0	0	0	0	5	0	0
18	1982	0	0	0	0	0	6	0	0
18	1983	0	0	0	0	0	7	0	0
18	1984	0	0	0	0	0	8	0	0

The first few lines and columns of `wagepan.tab` (the data set contains more variables than those listed above)

7.3 Suggested exercise

1. Estimate a linear model on `lwage` (log of hourly wage) without covariates, obtain the log likelihood, parameter estimates and standard errors. These results are given as the 1st part (homogeneous model) of the output that is obtained by estimating the random effects model. This is done in Task 2.

2. Allow for the person identifier (`nr`) random effect, use adaptive quadrature with mass 12. Is this random effect significant?

3. Add the covariates (`educ`, `black`, `hisp`, `exper`, `expersq`, `married`, `union`, factor(`year`). How does the magnitude of the `scale` parameter for person identifier random effects change?

4. Create interaction effects between the factor (`year`) indicators (`d81,...,d87`) and `educ`, add these effects to the previous model, do the returns to education vary with year? What do the results show?

7.4 References

Rabe-Hesketh, S., and Skrondal, A., (2005), Multilevel and Longitudinal Modelling using Stata, Stata Press, Stata Corp, College Station, Texas.

Vella, F., and Verbeek, M., (1998), Whose wages do unions raise? A dynamic model of unionism and wage rate determination for young men. Journal of Applied Econometrics, 13, 163-183.

Wooldridge, J. M., (2002), Econometric Analysis of Cross Section and Panel Data, MIT Press, Cambridge, MA.

8 Exercise L3. Linear Growth Model of log of Unemployment Claims

Papke (1994) analysed data from 1980 to 1988 to establish the effectiveness of Indiana's enterprise zone programme. This programme provided tax credits for cities with high poverty and unemployment levels. Papke (1994) was trying to establish if those cities in enterprise zones had lower unemployment claims. The same data were used by Rabe-Hesketh and Skrondal (2005, exercise 3.5).

8.1 Data description for `ezunem2.tab`

Number of observations (rows): 198
Number of level-2 cases: 22

8.2 Variables

`city`: city identifier (1,2,. . . ,22)
`year`: calendar year (1980,1981,. . . ,1988)
`uclms`: number of unemployment claims
`t`: linear time trend
`ez`: 1 if the city is in the enterprise zone, 0 otherwise
`d8m`: 1 if year is 198m, 0 otherwise, m=1,2,3,4,5,6,7,8
`cm`: 1 if city=m, 0 otherwise (m=1,2,. . . ,22)

city	year	uclms	t	ez	d81	d82	d83	d84	d85	d86	d87	d88	c1	c2
1	1980	166746	1	0	0	0	0	0	0	0	0	0	1	0
1	1981	83561	2	0	1	0	0	0	0	0	0	0	1	0
1	1982	158146	3	0	0	1	0	0	0	0	0	0	1	0
1	1983	83572	4	0	0	0	1	0	0	0	0	0	1	0
1	1984	45949	5	1	0	0	0	1	0	0	0	0	1	0
1	1985	48848	6	1	0	0	0	0	1	0	0	0	1	0
1	1986	46570	7	1	0	0	0	0	0	1	0	0	1	0
1	1987	47205	8	1	0	0	0	0	0	0	1	0	1	0
1	1988	37953	9	1	0	0	0	0	0	0	0	1	1	0
2	1980	115279	1	0	0	0	0	0	0	0	0	0	0	1
2	1981	78278	2	0	1	0	0	0	0	0	0	0	0	1
2	1982	126389	3	0	0	1	0	0	0	0	0	0	0	1
2	1983	79666	4	0	0	0	1	0	0	0	0	0	0	1
2	1984	41376	5	0	0	0	0	1	0	0	0	0	0	1
2	1985	53905	6	0	0	0	0	0	1	0	0	0	0	1

Some of the lines and columns of `ezunem2.tab` (the data set contains variables not used in this exercise)

8.3 Suggested exercise

1. Estimate a linear model on the log of number of unemployment claims (`luclms`) without covariates, obtain the log likelihood, parameter estimates and standard errors. These results are given as the 1st part (homogeneous model) of the output that is obtained by estimating the random effects model. This is done in Task 2.

2. Allow for the city identifier (`city`) random effect (use adaptive quadrature with mass 12). Is this random effect significant?

3. Add the binary `ez` effect. How does the magnitude of the `scale` parameter estimate for the city random effect change? Is the enterprise zone effect significant in this model?

4. Add the linear time effect (`t`). How does the magnitude of the city specific random effect change?

5. Interpret your preferred model, does `ez` have an effect on the response log(`uclms`)?

8.4 References

Papke, L. E., (1994), Tax policy and urban development: Evidence from the StateplaceIndiana enterprise zone program, Journal of Public Economics, 54, 37-49.

Rabe-Hesketh, S., and Skrondal, A., (2005), Multilevel and Longitudinal Modelling using Stata, Stata Press, Stata Corp, College Station, Texas.

9 Exercise L4. Binary Model of Trade Union Membership

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).

9.1 Data description for `wagepan.tab`

Number of observations (rows): 4360
Number of level-2 cases: 545

9.2 Variables

`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

nr	year	agric	black	bus	construc	ent	exper	fin	hisp
13	1980	0	0	1	0	0	1	0	0
13	1981	0	0	0	0	0	2	0	0
13	1982	0	0	1	0	0	3	0	0
13	1983	0	0	1	0	0	4	0	0
13	1984	0	0	0	0	0	5	0	0
13	1985	0	0	1	0	0	6	0	0
13	1986	0	0	1	0	0	7	0	0
13	1987	0	0	1	0	0	8	0	0
17	1980	0	0	0	0	0	4	0	0
17	1981	0	0	0	0	0	5	0	0
17	1982	0	0	0	0	0	6	0	0
17	1983	0	0	0	0	0	7	0	0
17	1984	0	0	0	0	0	8	0	0
17	1985	0	0	0	1	0	9	0	0
17	1986	0	0	0	1	0	10	0	0
17	1987	0	0	0	1	0	11	0	0
18	1980	0	0	0	0	0	4	0	0
18	1981	0	0	0	0	0	5	0	0
18	1982	0	0	0	0	0	6	0	0
18	1983	0	0	0	0	0	7	0	0
18	1984	0	0	0	0	0	8	0	0

The first few rows and columns of `wagepan.tab` (the data set contains other variables not used in this exercise)

9.3 Suggested exercise

1. Estimate a logit model for trade union membership (`union`), without covariates, obtain the log likelihood, parameter estimates and standard errors. These results are given as the 1st part (homogeneous model) of the output that is obtained by estimating the random effects model. This is done in Task 2.

2. Allow for the respondent identifier (`nr`) random effect, use adaptive quadrature. Is this random effect significant? How many quadrature points should we use to estimate this model?

3. Add the explanatory variables `black`, `hisp`, `exper`, `educ`, `poorhlth` and `married`. How does the magnitude of the `nr` random effect change? Are any of these individual characteristics significant in this model? Do the results make intuitive sense?

4. Add the contextual explanatory variables `rur, nrthcen, nrtheast, south`. How does the magnitude of the individual specific random effects coefficient change? Are any of the contextual variables significant in this model? Do the new results make intuitive sense?

5. Add the indicator variables for year. Are any of the year indicator variables significant in this model? Do the new results make intuitive sense?
6. Include interaction effects between `rur` and `nrthcen, nrtheast`, `south` and add them to the model. Are any of these new effects significant?
7. How can the final model be simplified?
8. Interpret your preferred model.

9.4 References

Rabe-Hesketh, S., and Skrondal, A., (2005), Multilevel and Longitudinal Modelling using Stata, Stata Press, Stata Corp, College Station, Texas.

Vella, F., and Verbeek, M., (1998), Whose wages do unions raise? A dynamic model of unionism and wage rate determination for young men. Journal of Applied Econometrics, 13, 163-183.

Wooldridge, J. M., (2002), Econometric Analysis of Cross Section and Panel Data, MIT Press, Cambridge, MA.

10 Exercise L5. Ordered Response Model of Attitudes to Abortion

Wiggins et al (1991) studied attitudes to abortion using a subset of the data from the British Social Attitudes (BSA) Survey. The BSA Survey is a multi-stage clustered random sample of adults (aged 18 and over) living in private households in Britain. The data are clustered by district.

A subset of individuals, from the 1983 BSA survey, were followed each year until 1986. The subset of the data we use here was used by Rabe-Hesketh and Skrondal (2005, exercise 5.5) for modelling the ordinal response strength of support for legalising abortion. The data are limited to the respondents who provided valid values for all 4 years of follow up. In this exercise we ignore any of the complications that may be caused by dropout from the follow up. The strength of support each year was judged by combining the responses (yes/no) on 7 different circumstances in which abortion should be legal. The questions relate to circumstances such as "The woman became pregnant as a result of rape", and "The woman decides on her own that she does not wish to have a child". Differences in magnitude of circumstances outside the woman's control are ignored and the respondent's total score is obtained by adding up the responses on the 7 different questions.

10.1 Data description for `abortion2.tab`

Number of observations (rows): 1056
Number of level-2 cases: 246

10.2 Variables

`district`: district identifier
`person`: respondent/individual identifier
`year`: year (1,2,3,4)
`score`: the number of questions (circumstances) to which the respondent answered yes
`age`: respondent's age in years
`male`: 1 if respondent is male, 0 otherwise
`nscore`: ordered response of attitude to abortion, for coding see below
`dr2`: 1 if the respondent's religion is protestant (catholic is the reference category), 0 otherwise
`dr3`: 1 if the respondent's religion is other religion, 0 otherwise
`dr4`: 1 if the respondent's religion is agnostic, 0 otherwise
`dp2`: 1 if the respondent votes labour (conservative is the reference category), 0 otherwise,
`dp3`: 1 if the respondent votes liberal, 0 otherwise
`dp4`: 1 if the respondent votes other, 0 otherwise
`dp5`: 1 if the respondent votes none, 0 otherwise
`dc2`: 1 if the respondent's self assessed social class is middle (upper is the reference category), 0 otherwise
`dc3`: 1 if the respondent's self assessed social class is lower, 0 otherwise

Coding of `nscore`

`nscore` = 1 if score=0,1,2 (as the values 0,1,2 for score are rare)
`nscore` = 2 if score =3
`nscore` = 3 if score =4
`nscore` = 4 if score =5
`nscore` = 5 ff score =6

district	person	year	score	age	male	nscore	dr2	dr3	dr4	dp2	dp3	dp4	dp5	dc2	dc3
4	39	1	3	49	1	2	0	0	1	0	1	0	0	0	1
4	39	4	3	49	1	2	0	0	1	0	1	0	0	0	1
4	39	2	7	49	1	6	0	0	1	0	1	0	0	0	1
4	39	3	3	49	1	2	0	0	1	0	1	0	0	0	1
4	46	2	3	50	0	2	0	1	0	0	0	0	0	1	0
4	46	1	3	50	0	2	0	1	0	0	0	0	0	1	0
4	46	3	7	50	0	6	0	1	0	0	0	0	0	1	0
4	46	4	7	50	0	6	0	1	0	0	0	0	0	1	0
4	48	4	4	51	0	3	1	0	0	1	0	0	0	0	1
4	48	2	4	51	0	3	1	0	0	1	0	0	0	0	1
4	48	3	3	51	0	2	1	0	0	1	0	0	0	0	1
4	48	1	6	51	0	5	1	0	0	1	0	0	0	0	1
4	55	4	7	21	1	6	0	0	1	1	0	0	0	1	0
4	55	2	7	21	1	6	0	0	1	1	0	0	0	0	1
4	55	3	6	21	1	5	0	0	1	1	0	0	0	0	1
4	55	1	6	21	1	5	0	0	1	0	0	0	0	0	1
4	56	1	7	27	1	6	0	0	0	1	0	0	0	0	1
4	56	3	7	27	1	6	0	0	0	1	0	0	0	1	0
4	56	2	5	27	1	4	0	0	0	1	0	0	0	1	0
4	56	4	7	27	1	6	0	0	0	1	0	0	0	1	0
4	60	2	3	72	0	2	1	0	0	0	0	0	0	0	0
4	60	3	5	72	0	4	1	0	0	0	0	0	0	0	0

The first few lines of `abortion2.tab`

10.3 Suggested exercise

1. Estimate an ordered logit model to `nscore`, without covariates, obtain the log likelihood, parameter estimates and standard errors. These results are given as the 1st part (homogeneous model) of the output that is obtained by estimating the random effects model. This is done in Task 2.

2. Allow for the person identifier (`person`) random effect, is this random effect significant? How many adaptive quadrature points should we use to estimate this model?

3. Add the explanatory variables `male`, `age` and the three sets of dummy variables (`dr, dp, dc`). How does the magnitude of the person random effect change? Are any of these individual characteristics significant in this model? Do the results make intuitive sense?

4. Repeat parts (2), (3) using `district` as the level-2 random effect, to do this you will need to use a version of the data set sorted by `district,` this has been done for you in `abortion3.tab`.

5. Does the significance of the explanatory variables change? Do the results make intuitive sense?

6. Interpret your preferred model. Can your preferred model be simplified?

7. Are there any interaction effects you would like to try to add to this model? Why?

10.4 References

Rabe-Hesketh, S., and Skrondal, A., (2005), Multilevel and Longitudinal Modelling using Stata, Stata Press, Stata Corp, College Station, Texas.

Wiggins, R. D., Ashworh, K., O'Muircheartaigh, C. A., Galbraith, J. J., (1991), Multilevel analysis of attitudes to abortion, Journal of the Royal Statistical Society, Series D, 40, 225-234.

11 Exercise L6. Ordered Response Model of Respiratory Status

Koch et al (1989) analysed the clinical trial data from 2 centres that compared two groups for respiratory illness. Eligible patients were randomised to treatment or placebo groups at each centre. The respiratory status (ordered response {0: terrible; 1: poor; 2: fair; 3: good; 4: excellent}) of each patient prior to randomisation and at 4 later visits to the clinic was determined.

The number of young patients in the sample is 110. The version of the data set `respiratory2.tab` we use was also used by Rabe-Hesketh and Skrondal (2005, exercise 5.1).

11.1 Data description for `respiratory2.tab`

Number of observations (rows): 555
Number of level-2 cases: 110

11.2 Variables

`center`: Centre (1,2)
`drug`: 1 if patient was allocated to the treatment group, 0 if placebo
`male`: 1 if patient was male, 0 otherwise
`age`: patient's age
`bl`: patient's respiratory status prior to randomisation
`v1`: patient's respiratory status at visit 1
`v2`: patient's respiratory status at visit 2
`v3`: patient's respiratory status at visit 3
`v4`: patient's respiratory status at visit 4
`patient`: Patient identifier (1,2,...,110)
`status`: the stacked versions of `bl` and `vt`, with 1 added to each value
`r1`: 1 if this is the response for `bl` (pre randomisation), 0 otherwise
`r2`: 1 if this is the response for visit 1, 0 otherwise
`r3`: 1 if this is the response for visit 2, 0 otherwise
`r4`: 1 if this is the response for visit 3, 0 otherwise
`r5`: 1 if this is the response for visit 4, 0 otherwise
`bld`: 1 if this is the pre randomisation response, 0 otherwise
`trend`: 0 or visit (1,2,3,4)
`base`: respiratory response at baseline

The data are sorted by `patient` within `center`.

ij	r	center	drug	male	age	bl	v1	v2	v3	v4	patient	status	r1	r2	r3	r4	r5	bld	trend	base
1	1	1	1	0	32	1	2	2	4	2	1	2	1	0	0	0	0	1	0	0
1	2	1	1	0	32	1	2	2	4	2	1	3	0	1	0	0	0	0	1	1
1	3	1	1	0	32	1	2	2	4	2	1	3	0	0	1	0	0	0	2	1
1	4	1	1	0	32	1	2	2	4	2	1	5	0	0	0	1	0	0	3	1
1	5	1	1	0	32	1	2	2	4	2	1	3	0	0	0	0	1	0	4	1
2	1	1	1	0	47	2	2	3	4	4	2	3	1	0	0	0	0	1	0	0
2	2	1	1	0	47	2	2	3	4	4	2	3	0	1	0	0	0	0	1	2
2	3	1	1	0	47	2	2	3	4	4	2	4	0	0	1	0	0	0	2	2
2	4	1	1	0	47	2	2	3	4	4	2	5	0	0	0	1	0	0	3	2
2	5	1	1	0	47	2	2	3	4	4	2	5	0	0	0	0	1	0	4	2
3	1	1	1	1	11	4	4	4	4	2	3	5	1	0	0	0	0	1	0	0
3	2	1	1	1	11	4	4	4	4	2	3	5	0	1	0	0	0	0	1	4
3	3	1	1	1	11	4	4	4	4	2	3	5	0	0	1	0	0	0	2	4
3	4	1	1	1	11	4	4	4	4	2	3	5	0	0	0	1	0	0	3	4
3	5	1	1	1	11	4	4	4	4	2	3	3	0	0	0	0	1	0	4	4
4	1	1	1	1	14	2	3	3	3	2	4	3	1	0	0	0	0	1	0	0
4	2	1	1	1	14	2	3	3	3	2	4	4	0	1	0	0	0	0	1	2
4	3	1	1	1	14	2	3	3	3	2	4	4	0	0	1	0	0	0	2	2

The first few lines of `respiratory2.tab`

11.3 Suggested exercise

1. Estimate an ordered logit model for status without any covariates, obtain the log likelihood, parameter estimates and standard errors. These results are given as the 1st part (homogeneous model) of the output that is obtained by estimating the random effects model. This is done in Task 2.

2. Estimate the ordered logit model for status, allowing for the patient random effect. Are the random patient effects significant? How many adaptive quadrature points should we use to estimate this model?

3. Re-estimate the model allowing for `drug`, `male`, `age` and `base`. How does the magnitude of the patient random effect change? Are any of these explanatory variables significant in this model? Do the results make intuitive sense?

4. Add the linear trend variable to the model, then add an interaction between `trend` and `drug`. Does the impact of treatment vary with visit?

11.4 References

Koch, G. G., Car, G. J., Amara, A., Stokes, M. E., and Uryniak, T. J., (1989), Categorical data analysis. In StateBerry, D., A., Statistical Methodology in the Pharmaceutical Sciences, pp 389-473, Marcel Dekker, New York.

Rabe-Hesketh, S., and Skrondal, A., (2005), Multilevel and Longitudinal Modelling using Stata, Stata Press, Stata Corp, College Station, Texas.

12 Exercise L8. Poisson Model of Epileptic Seizures

Thall and Vail (1990), Breslow and Clayton (1993) analysed longitudinal data on the number of epileptic seizures in successive intervals. The data were collected as part of a randomized controlled trial for the treatment of epilepsy which compared the treatment Progabide with a placebo. The data we use here was used by Rabe-Hesketh and Skrondal (2005, exercise 6.1). The data set `epilep.tab` have been stacked ready for analysis.

12.1 Data description for `epilep.tab`

Number of observations (rows): 236
Number of level-2 cases: 59

12.2 Variables

`subj`: Patient identifier
`y`: number of epileptic seizures over a two week period
`treat`: 1 if Progabide, 0 placebo
`visit`: visit time, coded as -0.3, -0.1, 0.1, 0.3
`v4`: 1 if the reponse relates to the 4^{th} visit, 0 otherwise (centered about its mean)
`lage`: logarithm of the patients age (centered about its mean)
`lbas`: logarithm of $\frac{1}{4}$ of the number of seizures in the 8 weeks preceding the trial, (centred about its mean)
`lbas.trt`: interaction between `lbas` and `treat` (centered about its mean)
`cons`: 1 for all observations

subj	y	treat	visit	v4	lage	lbas	lbas_trt	cons
1	5	0	-0.30	-0.25	0.11	-0.76	-0.95	1
1	3	0	-0.10	-0.25	0.11	-0.76	-0.95	1
1	3	0	0.10	-0.25	0.11	-0.76	-0.95	1
1	3	0	0.30	0.75	0.11	-0.76	-0.95	1
2	3	0	-0.30	-0.25	0.08	-0.76	-0.95	1
2	5	0	-0.10	-0.25	0.08	-0.76	-0.95	1
2	3	0	0.10	-0.25	0.08	-0.76	-0.95	1
2	3	0	0.30	0.75	0.08	-0.76	-0.95	1
3	2	0	-0.30	-0.25	-0.10	-1.36	-0.95	1
3	4	0	-0.10	-0.25	-0.10	-1.36	-0.95	1
3	0	0	0.10	-0.25	-0.10	-1.36	-0.95	1
3	5	0	0.30	0.75	-0.10	-1.36	-0.95	1
4	4	0	-0.30	-0.25	0.26	-1.07	-0.95	1
4	4	0	-0.10	-0.25	0.26	-1.07	-0.95	1
4	1	0	0.10	-0.25	0.26	-1.07	-0.95	1
4	4	0	0.30	0.75	0.26	-1.07	-0.95	1
5	7	0	-0.30	-0.25	-0.23	1.04	-0.95	1

The first few lines of `epilep.tab`

12.3 Suggested exercise

1. Estimate a Poisson model for the response number of epileptic seizures (`y`) with a constant but without any random effects, obtain the log likelihood, parameter estimates and standard errors. These results are given as the 1st part (homogeneous model) of the output that is obtained by estimating the random effects model. This is done in Task 2.

2. Re-estimate model (1) allowing for the patient effect (`subj`) random effects. Are the patient random effects significant? Use adaptive quadrature with mass 12.

3. Re-estimate model (2) allowing for `lbas`, `treat`, `lbas.trt`, `lage`, `visit` . How does the magnitude of the patient random effect change? Are any of these explanatory variables significant in this model? Do the results make intuitive sense?

4. Re-estimate model (3) adding `v4`, in place of visit, which model would you prefer?

5. Interpret your results. Can your preferred model be simplified?

6. Are there any other interaction effects you would like to try in this model? Why?

12.4 References

Breslow, N.E. & Clayton, D.G., (1993), Approximate inference in generalized linear mixed models, J. Am. Statist. Ass., 88, 9-25.

Thall, P. F. & Vail, S. C., (1990), Some covariance models for longitudinal count data with overdispersion, Biometrics, 46, 657-671.

13 Exercise L9. Bivariate Linear Model of Expiratory Flow Rates

Bland and Altman (1986) report on a study to compare the standard Wright peak flow meter with the (then) new Mini Wright peak flow meter. The data that accompany this study (`pefr.tab`) contain the repeated measurements of peak expiratory flow rate (PEFR) obtained from a sample of 17 individuals. These subjects had their PFER measured twice using the new Mini Wright peak flow meter and twice using the Standard Wright peak flow meter. To avoid instrument effects being confounded with prior experience effects, the instruments were used in random order.

13.1 Data description for `pefr.tab`

Number of observations (rows): 34
Number of level-2 cases: 17

13.2 Variables

`id`: person identifier
`occasion`: occasion {1,2}
`wp`: Standard Wright meter PEFR
`wm`: Mini Wright meter PEFR

id	occasion	wp	wm
1	1	494	512
1	2	490	525
2	1	395	430
2	2	397	415
3	1	516	520
3	2	512	508
4	1	434	428
4	2	401	444
5	1	476	500
5	2	470	500
6	1	557	600
6	2	611	625
7	1	413	364
7	2	415	460
8	1	442	380
8	2	431	390
9	1	650	658

The first few rows of `pefr.tab`

13.3 Suggested exercise

13.3.1 Standard Wright Meter: data set `pefr.tab`

1. Estimate a linear model for the response wp with occasion 2 (`occ2`) as a binary indicator with an `id` random effect. Is `occ2` significant? Are the random person effects (`id`) significant? Use adaptive quadrature with mass 12 and set the starting value for `scale` to 110.

13.3.2 Mini Wright Meter: data set `pefr.tab`

2 Estimate a linear model for the response `wm` with occasion 2 (`occ2`) as a binary indicator with an `id` random effect. Is `occ2` significant? Are the random person effects (`id`) significant? Use adaptive quadrature with mass 12 and set the starting value for `scale` to 100.

13.3.3 Joint Model: data set `pefr.tab`

3 Estimate a joint model for `wp` and `wm` with `occ2` as a binary indicator in both linear predictors, use adaptive quadrature with 12 mass points for both dimensions. As this is a very small data set the likelihood is not well defined. Use the following starting values: 0.9 for `rho`, 20 for both values of `sigma`, 110 for the first `scale` and 110 for the second. What is the significance of the correlation between the random effects of each type of meter? How does the significance of the `occ2` effect change, relative to that obtained in Task 1 and 2?

4 On the basis of these results would you be prepared to replace the Standard Wright flow meter with the new Mini Wright Meter?

13.4 References

Bland, J. M., and Altman, D., G., (1986), Statistical methods for assessing agreement between two methods of clinical measurement, Lancet, 1, 307-310.

14 Exercise L10. Bivariate Model, Linear (Wages) and Binary (Trade Union Membership)

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 wages and for separately modelling trade union membership by Rabe-Hesketh and Skrondal (2005, exercises 2.7 and 4.7). We start by re-estimating the separate models for log(wages) and for trade union membership. We then estimate a joint model allowing trade union membership to be endogenous in the wage equation.

14.1 Data description for `wagepan.tab`

Number of observations (rows): 4360
Number of level-2 cases: 545

14.2 Variables

`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 otherwise
`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
`lwage`: log of hourly wage in US dollars
`d8m`: 1 if the year is 198m, 0 otherwise, m=1,...,7

nr	year	agric	black	bus	construc	ent	exper	fin	hisp
13	1980	0	0	1	0	0	1	0	0
13	1981	0	0	0	0	0	2	0	0
13	1982	0	0	1	0	0	3	0	0
13	1983	0	0	1	0	0	4	0	0
13	1984	0	0	0	0	0	5	0	0
13	1985	0	0	1	0	0	6	0	0
13	1986	0	0	1	0	0	7	0	0
13	1987	0	0	1	0	0	8	0	0
17	1980	0	0	0	0	0	4	0	0
17	1981	0	0	0	0	0	5	0	0
17	1982	0	0	0	0	0	6	0	0
17	1983	0	0	0	0	0	7	0	0
17	1984	0	0	0	0	0	8	0	0
17	1985	0	0	0	1	0	9	0	0
17	1986	0	0	0	1	0	10	0	0
17	1987	0	0	0	1	0	11	0	0
18	1980	0	0	0	0	0	4	0	0
18	1981	0	0	0	0	0	5	0	0
18	1982	0	0	0	0	0	6	0	0
18	1983	0	0	0	0	0	7	0	0
18	1984	0	0	0	0	0	8	0	0

The first few rows and columns of `wagepan.tab` (for the univariate models)

14.3 Suggested exercise

14.3.1 Univariate models

14.3.2 Wage equation: data `wagepan.tab`

1. Estimate a linear model for `lwage` (log of hourly wage) with the covariates (`educ`, `black`, `hisp`, `exper`, `expersq`, `married`, `union`) , with the data clustered over time for `nr` (respondent identifier) Is this random effect significant? Use adaptive quadrature, mass 12.

14.3.3 Trade union membership: data `wagepan.tab`

2 Estimate a logit model for trade union membership (`union`), with the covariates (`black`, `hisp`, `exper`, `educ`, `poorhlth`, `married`, `rur`, `nrthcen`, `nrtheast`, `south`). Use adaptive quadrature, mass 64. Use `case nr,` (respondent identifier). Is this random effect significant?

14.3.4 Joint model: data `wagpan.tab`

3 Using the model specifications for log(wages) and trade union membership you have just used, estimate a joint model of the determinants of

log(wages) and trade union membership. Use adaptive quadrature, mass 4 for the linear model and mass 64 for the binary response.

4 What is the magnitude and significance of the correlation between the random effects for log(wages) and union membership? How does the magnitude and significance of the direct effect of union in the wage equation change? What are the reasons for this? Have any other features of the models changed? What does this imply?

14.4 References

Rabe-Hesketh, S., and Skrondal, A., (2005), Multilevel and Longitudinal Modelling using Stata, Stata Press, Stata Corp, College Station, Texas.

Vella, F., and Verbeek, M., (1998), Whose wages do unions raise? A dynamic model of unionism and wage rate determination for young men. Journal of Applied Econometrics, 13, 163-183.

Wooldridge, J, M., (2002), Econometric Analysis of Cross Section and Panel Data, MIT Press, Cambridge, MA.

15 Exercise L11. Renewal Model of Angina Pectoris (Chest Pain)

Pickles and Crouchley (1994) analysed a sub set of the data from Danahy et al (1977) on the length of exercise time (seconds) required to induce angina pectoris in 21 heart patients. The subset they used was for the times to angina: just before oral administration of a dose of isosorbide dinitrate, one hour after and three hours after administration. Eleven of the 63 exercise times were censored due to patient fatigue. This censoring process was assumed to be independent of the frailty (random effects) for Angina. Pickles and Crouchley (1994) used a Positive Stable Law distribution for the frailty. This exercise will repeat their analysis using a lognormal distribution for the frailty (normal distribution for the random effects). In Pickles and Crouchley (1997) the exercise data was treated as continuous responses. Rather that treat the data as continuous, the data have been expanded so that each second of exercise time is a discrete interval of time (`angina.tab`). The duration of the current interval of exercise is measured from the start of that session of exercise. The exercise will explore whether the impact of dose declines with distance from the treatment, whether the duration effects also change with distance form treatment in a renewal model.

	Time		Dose		Time		Dose
0	1	3		0	1	3	
136	445+	393+	0.58	147	403	290	0.44
250	306	206	0.34	231	540+	370	0.49
215	232	258	0.24	224	432	291	0.31
235	248	298	0.37	152	733+	492	0.2
129	121	110	0.38	417	743+	566	0.24
425	580	613	0.32	213	250	150	0.38
441	504+	519+	0.41	490	559+	557+	0.27
208	264	210	0.37	406	651	624	0.51
154	110	123	0.37	229	327	280	0.24
89	145	172	0.53	265	565+	505+	0.51
250	230	264	0.24				

Note: + Observations censored by fatigue

A subset of the Angina data from Danahy et al (1977)

The subset of data from Danahy et al (1977), from the above table has been rearranged in discrete time intervals (seconds) for this exercise.

15.1 Data description for `angina.tab`

Number of observations: 20985
Number of level-2 cases: 21

15.2 Variables

`id`: patient identifier
`d`: time, collapsed to 1 = pre-dose and 2 = post-dose
`time`: 1 = pre-dose, 2 = 1 hour post-dose, 3 = 3 hours post-dose
`dose`: dosage
`t`: exercise time in seconds
`y`: response, 1 if observation censored by fatigue. 0 otherwise
`d1`: 1 if d = 1, 0 otherwise
`d2`: 1 if d = 2, 0 otherwise
`t1`: 1 if t = 1, 0 otherwise
`t2`: 1 if t = 2, 0 otherwise
`t3`: 1 if t = 3, 0 otherwise

id	d	time	dose	t	y	censored	d1	d2	t1	t2	t3
1	1	1	0.579999983	1	0	0	1	0	1	0	0
1	1	1	0.579999983	2	0	0	1	0	1	0	0
1	1	1	0.579999983	3	0	0	1	0	1	0	0
1	1	1	0.579999983	4	0	0	1	0	1	0	0
1	1	1	0.579999983	5	0	0	1	0	1	0	0
1	1	1	0.579999983	6	0	0	1	0	1	0	0
1	1	1	0.579999983	7	0	0	1	0	1	0	0
1	1	1	0.579999983	8	0	0	1	0	1	0	0
1	1	1	0.579999983	9	0	0	1	0	1	0	0
1	1	1	0.579999983	10	0	0	1	0	1	0	0
1	1	1	0.579999983	11	0	0	1	0	1	0	0
1	1	1	0.579999983	12	0	0	1	0	1	0	0
1	1	1	0.579999983	13	0	0	1	0	1	0	0
1	1	1	0.579999983	14	0	0	1	0	1	0	0
1	1	1	0.579999983	15	0	0	1	0	1	0	0
1	1	1	0.579999983	16	0	0	1	0	1	0	0
1	1	1	0.579999983	17	0	0	1	0	1	0	0
1	1	1	0.579999983	18	0	0	1	0	1	0	0

First few lines of `angina.tab` (discrete time version of the data from Danahy et al, 1977)

15.3 Suggested exercise

1. We are going to estimate various Weibull survival models on the renewal data by using (`logt`) as a covariate with the cloglog link. The 1st model is the homogeneous common baseline hazard model, i.e. with the same constant for each exercise time, the same parameter for `logt,` but with different coefficients on `dose` for the two treatment times, use interactions with the `t2` and `t3` dummy variables to set this model up. Obtain the log likelihood, parameter estimates and standard errors. These results are

given as the 1st part (homogeneous model) of the output that is obtained by estimating the random effects model. These model results can also can be obtained from sabreR by putting mass=1. There is no point putting `dose` in the linear predictor for the model of pre-treatment data.

2. The 2nd model allows for a different baseline hazard for each exercise session. Interact the `t2` and `t3` dummy variables with `logt,` add both the interaction effects and the `t2` and `t3` dummies to the model. Obtain the log likelihood, parameter estimates and standard errors. These results are given as the 1st part (homogeneous model) of the output that is obtained by estimating the random effects model. These model results can also can be obtained from sabreR by putting mass=1. Can the model be simplified? What does this result tell you?

3 Add a subject specific random effect (`id)` to the renewal model. Use adaptive quadrature with mass 24. How do the effects of `logt` and `dose` change, relative to the models estimated in Task 1 and 2?

3. What is your preferred model and why?

15.4 References

Danahy, D.J., Burwell, D.T., Aranow, W.S., Parkash, R., (1977), Sustained henodynamic and anti-anginal effect of high dose oral isosorbide dinitrate, Circulation, 55, 381-387.

Pickles A.R. and Crouchley, R., (1994), Generalizations and Applications of Frailty models for Survival and Event Data, Statistical Models in Medical Research, 3, 263-278.

16 Exercise L12. Bivariate Competing Risk Model of German Unemployment Data

The data for this exercise are for the time spent in unemployment with exits to two destinations: full time and part time reemployment. The data are from the German Socio Economic Panel (SOEP), www.diw.de/deutsch/sop. The data set (`unemployedR.tab`) contains spells of unemployment for 500 individuals. The observations or spells are clustered according to the identification number of the person. Time spent in the unemployment spell is measured in months. The spells which lasted more than 36 months have been censored at 36 months. The data is available from Cran, see http://cran.r-project.org/web/packages/CompetingRiskFrailty/index.html. The data form part of the example of the software developed by Kauermann and Khomski (2006a, b). The data for this exercise have been written out in discrete form using months.

16.1 Data description for `unemployedR.tab`

Number of observations (rows): 6070
Number of level-2 cases: 500

16.2 Variables

`id`: individual identifier
`t`: unemployment duration in months
`survival`: total length of unemployment spell in months
`full`: exit to full-time employment
`part`: exit to part-time employment
`nationality`: nationality (1 = German, 2 = foreign)
`gender`: gender (1 = male, 2 = female)
`age`: age (1 = 25 or younger, 2 = aged 26-50, 3 = older than 50)
`training`: training (1 = professional training, 2 = otherwise)
`university`: university (1 = no degree, 2 = degree)
`rowname`: row number from unexpanded data
`spell`: individual-level unemployment spell
`y`: 1 if exit to employment, 0 otherwise
`r`: risk variate (1 = full-time, 2 = part-time)
`r1`: 1 if r=1, 0 otherwise
`r2`: 1 if r=2, 0 otherwise
`id_spell`: combined individual-spell identifier
`age1`: 1 if age=1, 0 otherwise
`age2`: 1 if age=2, 0 otherwise
`age3`: 1 if age=3, 0 otherwsie

id	t	survival	full	part	nationality	gender	age	training	university	rowname	spell	y	r	r1	r2	id_spell	age1	age2	age3
916102	1	1	1	0	1	1	1	1	2	5954	1	1	1	1	0	9161021	1	0	0
916102	1	1	1	0	1	1	1	1	2	5954	1	0	2	0	1	9161021	1	0	0
916602	1	3	1	0	1	2	1	1	1	5955	1	0	1	1	0	9166021	1	0	0
916602	2	3	1	0	1	2	1	1	1	5955	1	0	1	1	0	9166021	1	0	0
916602	3	3	1	0	1	2	1	1	1	5955	1	1	1	1	0	9166021	1	0	0
916602	1	3	1	0	1	2	1	1	1	5955	1	0	2	0	1	9166021	1	0	0
916602	2	3	1	0	1	2	1	1	1	5955	1	0	2	0	1	9166021	1	0	0
916602	3	3	1	0	1	2	1	1	1	5955	1	0	2	0	1	9166021	1	0	0
916602	1	5	1	0	1	2	1	1	1	5956	2	0	1	1	0	9166022	1	0	0
916602	2	5	1	0	1	2	1	1	1	5956	2	0	1	1	0	9166022	1	0	0
916602	3	5	1	0	1	2	1	1	1	5956	2	0	1	1	0	9166022	1	0	0
916602	4	5	1	0	1	2	1	1	1	5956	2	0	1	1	0	9166022	1	0	0
916602	5	5	1	0	1	2	1	1	1	5956	2	1	1	1	0	9166022	1	0	0
916602	1	5	1	0	1	2	1	1	1	5956	2	0	2	0	1	9166022	1	0	0
916602	2	5	1	0	1	2	1	1	1	5956	2	0	2	0	1	9166022	1	0	0

First few lines of `unemployedR.tab`

16.3 Suggested exercise

1. Estimate a Weibull (`logt`), non random effects model, for the `r1=1` (full time job) and `r2=1` (part time job) exits from unemployment, use the covariates: `nationality, gender, age, training`, `university.` Obtain the log likelihood, parameter estimates and standard errors. These results are given as the 1st part (homogeneous model) of the output that is obtained by estimating the random effects model. This is done in Task 2.

2. Re-estimate the model from question 1 but allow each exit type to have an independent random effect for each failure type, use 32 point adaptive quadrature. Hint, use a bivariate model, but set `rho=0`. What do the results tell you?

3. Re-estimate the model from question 2 but allow for the correlation between the random effects of each failure type. How do the results change?

4. What is your preferred model and why?

16.4 References

Kauermann G. and Khomski P. (2006a), Additive two way hazards model with varying coefficients, in press.

Kauermann G. and Khomski P. (2006b), Full Time or Part Time Reemployment: A Competing Risk Model with Frailties and Smooth Effects using a Penalty based Approach, see http://www.wiwi.uni-bielefeld.de/~kauermann/research/Competing_Risk_Model.pdf.

17 Exercise 3LC1. Linear Model: Pupil Rating of School Managers (856 Pupils in 94 Schools)

This data set (`manager.tab`) was presented by Hox (2002) and contains the response '`scores`' given by each pupil on 6 questions on the nature of their school managers/directors, for a collection of schools. The data set also contains information on the director's gender, the type of the school, the pupil gender and year of the pupil. Hox (2002) presents the results for a 3-level linear model (without explanatory variables) in Hox (2002, Table 9.5). For details about the book see `http://www.geocities.com/joophox/mlbook/leabook.htm`

17.1 Data description for `manager.tab`

Number of observations: 4981
Number of level-2 cases ('`pupil`'): 856
Number of level-3 cases ('`school`'): 94

17.2 Variables

`id`: pupil identifier across all schools
`school`: school identifier
`pupil`: pupil identifier within each school (0,1,. . . 9)
`dirsex`: gender of school manager (1: F, 2: M)
`schtype`: school type (1=general (AVO), 2=professional (MBO &T), 3= day/evening)
`pupsex`: pupil gender (1= F, 2=M)
`item`: item (1,2,. . . ,6)
`cons`: constant
`class`: school year of pupil
`scores`: response by pupil to the item question.

id	school	pupil	dirsex	schtype	pupsex	item	cons	class	scores
1	6	0	2	2	1	1	1	2	4
1	6	0	2	2	1	2	1	2	4
1	6	0	2	2	1	3	1	2	3
1	6	0	2	2	1	4	1	2	2
1	6	0	2	2	1	5	1	2	2
1	6	0	2	2	1	6	1	2	3
2	6	1	2	2	1	1	1	2	1
2	6	1	2	2	1	2	1	2	1
2	6	1	2	2	1	3	1	2	1
2	6	1	2	2	1	4	1	2	1
2	6	1	2	2	1	5	1	2	3
2	6	1	2	2	1	6	1	2	2
3	6	2	2	2	1	1	1	2	4
3	6	2	2	2	1	2	1	2	4
3	6	2	2	2	1	3	1	2	4
3	6	2	2	2	1	4	1	2	2
3	6	2	2	2	1	5	1	2	1
3	6	2	2	2	1	6	1	2	2
4	6	3	2	2	1	1	1	2	3
4	6	3	2	2	1	2	1	2	3
4	6	3	2	2	1	3	1	2	3
4	6	3	2	2	1	4	1	2	2
4	6	3	2	2	1	5	1	2	2
4	6	3	2	2	1	6	1	2	3
5	6	4	2	2	1	1	1	2	4
5	6	4	2	2	1	2	1	2	4
5	6	4	2	2	1	3	1	2	4
5	6	4	2	2	1	4	1	2	3
5	6	4	2	2	1	5	1	2	2

The first few lines of `manager.tab`

17.3 Suggested exercise:

1. Estimate a linear model (without random effects) for the `scores` with the pupil- and school- level covariates `dirsex`, `schtype` and `pupsex`. These results are given as the 1st part (homogeneous model) of the output that is obtained by estimating the random effects model. This is done in Task 2.

2. Allow for the pupil identifier random effect (`id`), use adaptive quadrature with `mass`=12, in a 2-level model. Is this random effect significant?

3. Allow for both the pupil identifier random effect (`id`) and for the school random effect (`school`) in a 3-level model, use adaptive quadrature with mass 24 for both levels. Are both these random effects significant? Is this model a significant improvement over the model estimated in part 2 of this exercise?

4. Which covariates have a significant effect on the scores? How did your results change when you allowed for pupil-level (level 2) and then school-level (level 3) effects?

17.4 References

Hox, J., (2002), Multilevel Analysis Techniques and Applications, Lawrence Erlbaum Associates, London

18 Exercise 3LC2. Binary Response Model for the Tower of London tests (226 Individuals in 118 Families)

This data set (`towerl.tab`) is from Rabe-Hesketh and Skrondal (2005). Rabe-Hesketh, Touloupolou and Murray (2001) estimated a multilevel cognitive performance model on 3 groups: (1) subjects with schizophrenia; (2) subject's relatives and (3) control subjects. The Tower of London test was used to assess cognitive performance. The responses have a 3-level structure, i.e. occasion i for subject j in family k. The test was repeated at 3 different levels of difficulty. The binary response `dtlm` takes the value 1 if each test was completed in the minimum number of moves and 0 otherwise. The same data were used by Rabe-Hesketh and Skrondal (2005, exercise 7.2).

18.1 Data description for `towerl.tab`

Number of observations: 677
Number of level-2 cases (`id`: subject identifier): 226
Number of level-3 cases (`famnum`: family identifier): 118

18.2 Variables

`id`: subject identifier
`level`: level of difficulty of the Tower of London test
`famnum`: family identifier
`group`: group (1=controls, 2=relatives, 3=schizophrenics)
`age`: subject's age (years)
`dtlm`: 1 if respondent completed the task in the minimum number of moves, 0 otherwise

id	level	famnum	group	age	sex	tlm	tlpl	tlcpl	tlsub	tlcsub	occ	dtlm
1	-1	14	3	30	1	1.253	0.483	0.300	2.207	1.539	3	0
1	0	14	3	30	1	2.140	0.207	0.419	3.450	1.826	4	0
1	1	14	3	30	1	1.705	0.884	0.351	2.682	2.014	5	0
2	-1	18	3	29	1	1.253	0.466	0.378	1.479	1.206	3	0
2	0	18	3	29	1	2.788	0.295	0.077	4.053	1.258	4	0
2	1	18	3	29	1	2.565	0.239	0.262	3.118	1.575	5	0
3	-1	21	3	44	1	1.179	0.523	0.542	1.522	1.493	3	0
3	0	21	3	44	1	1.833	0.310	0.577	2.912	1.670	4	0
3	1	21	3	44	1	1.981	0.534	0.713	3.043	1.908	5	0
4	-1	19	3	34	2	1.099	0.658	0.610	1.379	1.230	3	1
4	0	19	3	34	2	1.504	0.879	0.582	2.727	1.486	4	0
4	1	19	3	34	2	1.749	0.871	0.531	2.453	1.848	5	0
5	-1	16	3	39	2	1.099	0.216	0.278	1.468	1.609	3	1
5	0	16	3	39	2	1.658	0.594	0.113	2.782	1.914	4	0
5	1	16	3	39	2	1.658	0.841	0.207	2.514	2.103	5	0
6	-1	5	3	42	1	1.179	0.495	1.898	2.215	2.052	3	0
6	0	5	3	42	1	2.225	0.699	1.923	3.928	2.366	4	0
6	1	5	3	42	1	2.015	1.115	1.026	3.469	2.467	5	0
7	-1	6	3	53	1	1.099	0.727	0.859	1.573	1.376	3	1
7	0	6	3	53	1	2.197	0.351	0.560	3.316	1.603	4	0
7	1	6	3	53	1	1.833	0.410	0.293	2.444	1.870	5	0
8	-1	15	3	23	1	1.099	0.860	0.285	1.504	1.303	3	1
8	0	15	3	23	1	1.910	0.454	0.207	2.740	1.558	4	0
8	1	15	3	23	1	2.110	0.579	0.315	2.956	1.712	5	0
9	-1	10	3	29	1	1.179	0.059	0.344	1.144	1.215	3	0
9	0	10	3	29	1	1.833	0.688	0.285	2.415	1.597	4	0
9	1	10	3	29	1	2.015	0.940	0.247	2.992	1.660	5	0
10	-1	10	3	27	1	1.099	0.190	-0.020	0.846	1.026	3	1

The first few lines of `towerl.tab`

18.3 Suggested exercise

1. Estimate a logit model (without random effects) for the binary response `dtlm` with the covariate `level`, and dummy variables for `group`=2 and `group`=3. These results are given as the 1st part (homogeneous model) of the output that is obtained by estimating the random effects model. This is done in Task 2.

2. Allow for the level-2 subject random effect (`id`), use adaptive quadrature with `mass 12`. Is this random effect significant?

3. Allow for both the level-2 subject random effect (`id`), and for the level-3 family random effects (`famnum`), use adaptive quadrature with `mass 12.` Are both these random effects significant? Is this model a significant improvement over the model estimated in part 2 of this exercise?

4. How did your results on `group=2` and `group=3` change when you allowed for `subject` (level 2) and then `family` (level 3) effects?

18.4 References

Rabe-Hesketh, S., Toulopoulou, T. and Murray, R. (2001). Multilevel modeling of cognitive function in schizophrenic patients and their first degree relatives. Multivariate Behavioral Research 36, 279-298.

Rabe-Hesketh, S., and Skrondal, A., (2005), Multilevel and Longitudinal Modelling using Stata, Stata Press, Stata Corp, College Station, Texas.

19 Exercise 3LC3. Binary Response Model of the Guatemalan Immunisation of Children (1595 Mothers in 161 Communities)

This exercise uses the Rodríguez and Goldman (2001) data on Guatemalan families, decisions whether or not to immunize their children. The survey was conducted in 1987, in order to establish the effectiveness of the Guatemalan government's campaign to immunize children against major childhood diseases. The questionnaire contains information on the immunization status of alive children born in the previous 5 years. If the child was more than 2 years old at the time of the interview they were old enough to be immunized during the 1986 campaign. The data set contains the binary response `immun` which represents whether the child was immunized (1 yes, 0 otherwise) for child i in family j (level 2), within community k (level 3). The same data (`guatemala_immun.tab`) were used by Rabe-Hesketh and Skrondal (2005, section 7.5).

19.1 Data description for `guatemala_immun.tab`

Number of observations: 2159
Number of level-2 cases (`mom`: identifier for mothers): 1595
Number of level-3 cases (`cluster`: identifier for communities): 161

19.2 Variables

`kid`: child identifier
`mom`: identifier for mothers
`cluster`: identifier for communities
`immun`: 1 if the child was immunized, 0 otherwise
`kid2p`: 1 if child aged 2-3 years, 0 otherwise
`mom25p`: 1 if mother aged 25+ years, 0 otherwise
`order23`: 1 if birth order 2-3, 0 otherwise
`order46`: 1 if birth order 4-6, 0 otherwise
`order7p`: 1 if birth order 7+, 0 otherwise
`indnospa`: 1 if indigenous and speaks no Spanish, 0 otherwise
`inspa`: 1 if indigenous and speaks Spanish, 0 otherwise
`momedpri`: 1 if mother's education primary, 0 otherwise
`momedsec`: 1 if mother's education secondary+, 0 otherwise
`husedppri`: 1 if husband's education primary, 0 otherwise
`husedsec`: 1 if husband's education secondary+, 0 otherwise
`huseddk`: 1 if husband's education missing, 0 otherwise
`momwork`: 1 if mother working , 0 otherwise
`rural`: 1 if identifier for a rural community, 0 otherwise
`pcind81`: proportion indigenous in 1981

kid	mom	cluster	immun	kid2p	mom25p	order23	order46	order7p	indNoSpa	indSpa	momEdPri	momEdSec	husEdPri	husEdSec	husEdDK	momWork	rural	pcInd81
2	2	1	1	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0.11
269	185	36	0	1	0	1	0	0	0	0	1	0	1	0	0	1	0	0.04
272	186	36	0	1	0	0	0	0	0	0	1	0	0	1	0	1	0	0.04
273	187	36	0	1	0	1	0	0	0	0	1	0	1	0	0	1	0	0.04
274	188	36	0	1	1	0	1	0	0	0	1	0	0	0	1	1	0	0.04
275	188	36	1	1	0	1	0	0	0	0	1	0	0	0	1	1	0	0.04
276	189	36	1	1	0	1	0	0	0	0	0	1	1	0	0	1	0	0.04
277	190	36	1	0	1	0	0	1	0	0	1	0	1	0	0	1	0	0.04
278	190	36	1	1	1	0	1	0	0	0	1	0	1	0	0	1	0	0.04
280	191	36	1	1	1	0	1	0	0	0	1	0	0	1	0	1	0	0.04
281	192	36	0	0	0	1	0	0	0	0	0	1	0	1	0	1	0	0.04
282	192	36	1	1	0	0	0	0	0	0	0	1	0	1	0	1	0	0.04
299	204	38	1	1	1	0	1	0	0	0	1	0	1	0	0	1	0	0.04
300	205	38	1	1	1	0	0	0	0	0	1	0	0	0	1	1	0	0.04
301	206	38	1	1	1	0	1	0	0	0	1	0	0	1	0	1	0	0.04
358	245	45	0	1	1	1	0	0	0	0	1	0	1	0	0	1	0	0.01
359	245	45	0	1	0	1	0	0	0	0	1	0	1	0	0	1	0	0.01
365	248	45	1	1	1	1	0	0	0	0	1	0	1	0	0	1	0	0.01
366	249	45	0	1	1	0	1	0	0	0	0	0	1	0	0	0	0	0.01
368	250	45	1	1	0	0	1	0	0	0	1	0	1	0	0	0	0	0.01
369	250	45	0	1	0	1	0	0	0	0	1	0	1	0	0	0	0	0.01
371	251	45	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0.01
372	252	45	0	1	1	0	0	1	0	0	1	0	1	0	0	0	0	0.01
373	253	45	0	0	1	0	0	1	0	0	1	0	1	0	0	1	0	0.01
374	253	45	0	1	1	0	1	0	0	0	1	0	1	0	0	1	0	0.01
375	254	45	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0.01
376	254	45	1	1	1	0	1	0	0	0	0	0	1	0	0	0	0	0.01
377	255	45	1	1	0	1	0	0	0	0	0	1	0	1	0	1	0	0.01

The first few lines of `guatemala_immun.tab`

19.3 Suggested exercise

1. Estimate a logit model (without random effects) for the binary response `immun` with a constant and the covariates `kid2p, mom25p, order23, order46, order7p, indnospa, indspa, momedpri, momedsec, husedpri, husedsec, huseddk, momwork, rural` and `pcind81`. These results are given as the 1st part (homogeneous model) of the output that is obtained by estimating the random effects model. This is done in Task 2.

2. Allow for the family random effect (`mom`), use adaptive quadraure with `mass 24`. Is this random effect significant?

3. Allow for both the level 2 family random effect (`mom`) and for the level 3 community random effects (`cluster`), use adaptive quadraure with `mass 32` for both levels. Are both these random effects significant? Is this model a significant improvement over the model estimated in part 2 of this exercise?

4. How did your covariate inference change when you allowed for mom-level (level 2) and then community-level (`cluster`, level 3) effects?

19.4 References

Rodriguez, G., and Goldman, N., (2001), Improved estimation procedures for multilevel models with binary response: a case study. Journal of the Royal Statistical Society, A 164, 339–355.

Rabe-Hesketh, S., and Skrondal, A., (2005), Multilevel and Longitudinal Modelling using Stata, Stata Press, Stata Corp, College Station, Texas.

20 Exercise 3LC4. Poisson Model of Skin Cancer Deaths (78 Regions in 9 Nations)

This exercise uses the Langford et al (1998) data from the Atlas of Cancer Mortality in the European Economic Community (Smans et al, 1992). Data were collected on male malignant melanoma deaths over the period 1975 to 1981 for the UK, Ireland, Italy, Germany, the Netherlands and for 1971-1980 for other EEC countries. Interest focuses on establishing the role of ultraviolet (`uv`) light exposure to malignant melanoma deaths. The data set (`deaths.tab`) contains the number of deaths by year in county i (level 1) within `region` j (level 2), within `nation` k (level 3). The same data were used by Rabe-Hesketh and Skrondal (2005, exercises 6.4, 7.5).

20.1 Data description for `deaths.tab`

Number of observations: 354
Number of level-2 cases (`region`: region identifier (EEC level-I areas)): 78
Number of level-3 cases (`nation`: nation identifier): 9

20.2 Variables

`nation`: nation identifier
`region`: region identifier
`county`: county identifier
`deaths`: number of male deaths due to malignant melanoma (skin cancer) during 1971-1980
`expected`: number of expected deaths
`uvb`: measure of the UVB dose reaching the earth's surface in each county and centered around its mean
`mr`: mortality rate

nation	region	county	deaths	expected	uvb	mr
1	1	1	79	51.222	-2.906	154.231
1	2	2	80	79.956	-3.207	100.055
1	2	3	51	46.517	-2.804	109.638
1	2	4	43	55.053	-3.007	78.107
1	2	5	89	67.758	-3.007	131.350
1	2	6	19	35.976	-3.418	52.813
1	3	7	19	13.280	-2.667	143.072
1	3	8	15	66.558	-2.667	22.537
1	3	9	33	50.969	-3.122	64.745
1	3	10	9	11.171	-2.485	80.566
1	3	11	12	19.683	-2.529	60.966
2	4	12	156	108.040	-1.138	144.391
2	4	13	110	73.692	-1.398	149.270
2	4	14	77	57.098	-0.439	134.856
2	4	15	56	46.622	-1.025	120.115
2	5	16	220	112.610	-0.503	195.365
2	5	17	46	30.334	-1.461	151.645
2	5	18	47	29.973	-1.896	156.808
2	5	19	50	32.027	-2.554	156.118
2	5	20	90	46.521	-1.967	193.461
2	5	21	62	36.990	-2.344	167.613
2	5	22	85	46.942	-0.658	181.075
2	6	23	141	55.383	-3.884	254.591
2	7	24	38	21.304	-4.459	178.370
2	8	25	121	50.229	-4.858	240.897
2	9	26	218	136.080	-2.603	160.200
2	9	27	50	36.712	-3.535	136.195
2	10	28	97	50.625	-4.025	191.605

The first few lines of `deaths.tab`

20.3 Suggested exercise

1. Estimate a Poisson model (without random effects) for the number of deaths (`deaths`) with the covariate `uvb`. Use log `expected` deaths as an offset. These results are given as the 1st part (homogeneous model) of the output that is obtained by estimating the random effects model. This is done in Task 2.

You will need accurate arithmetic for the following questions.

2 Allow for the level-2 region random effect (`region`), use adaptive quadrature with `mass 12`. Is this random effect significant?

3 Re-estimate the model with the level-2 random effect (`region`) and with `nation` as a level-3 random effect (`nation`). Use adaptive quadrature with `mass 96` for both levels. Are both these random effects significant?

4 How did your inference for the estimate of `uvb` change when you allowed for region-level (level 2) and then nation-level (level 3) effects?

20.4 References

Langford, I.H., Bentham, G., McDonald, A., (1998) Multilevel modelling of geographically aggregated health data: a case study on malignant melanoma mortality and UV exposure in the European Community, Statistics in Medicine, 17, pp 41-58.

Rabe-Hesketh, S., and Skrondal, A., (2005), Multilevel and Longitudinal Modelling using Stata, Stata Press, Stata Corp, College Station, Texas.

Smans, M., Muir, C.S., Boyle, P., (1992), Atlas of Cancer Mortality in the European Economic Community, Lyon, France: IARC Scientific Publications.

21 Exercise 3LC5. Event History Cloglog Link Model of Time to Fill Vacancies (1736 Vacancies in 515 Firms)

This is a study of the length of time (level 1, observed at the weekly level) needed to fill vacancies (level 2) by employers (level 3) in the vacancy data sub set `vwks_30k.tab`. We estimate a stock model of the duration of the vacancy; in addition to the firm's characteristics and those of the vacancy, we use covariates which represent the stock of the labour market at the current duration, i.e. the total number of job-seekers (logged) and the total number of vacancies (logged) in the local labour market.

21.1 Data description for `vwks4_30k.tab`

Number of observations: 28791 (weeks)
Number of level-2 cases (`vacref:` identifier for vacancy): 1736
Number of level-3 cases (`empref:` identifier for firm): 515

21.2 Variables

`match`: 1 if vacancy filled in a particular week, 0 otherwise
`nonman`: 1 if a non-manual vacancy, 0 otherwise
`written`: 1 if vacancy required a written method of application, 0 otherwise
`size`: firm size of the vacancy
`wage`: log wage of the vacancy
`vacref`: vacancy reference (a number)
`grade`: grade required by the vacancy
`empref`: employer reference (a number)
`dayrel`: 1 if day release available to the post, 0 otherwise
`t`: vacancy duration (see below)
`loguu`: log of stock of job-seekers in the local labour market
`logvv`: log of stock of vacancies in the local labour market

The covariate (`t`) for the baseline hazard is defined as follows:

`t=` 1 for week 1
`t=` 2 for week 2
`t=` 3 for weeks 3-4
`t=` 4 for weeks 5-6
`t=` 5 for weeks 7-8
`t=` 6 for weeks 9-13
`t=` 7 for weeks 14-26
`t=` 8 for weeks 27-39
`t=` 9 for weeks 40-52
`t=` 10 for weeks 53+

match	nonman	written	size	wage	vacref	grade	empref	dayrel	t	loguu	logvv
0	0	0	2	1.82	17500	1	1	0	1	7.05	4.63
0	0	0	2	1.51	18776	2	1	0	1	7.56	5.08
0	0	0	2	1.51	18776	2	1	0	2	7.88	5.10
0	0	0	2	1.51	18776	2	1	0	3	7.93	5.15
0	0	0	2	1.51	18776	2	1	0	3	7.91	5.19
0	0	0	2	1.97	20017	1	1	0	1	7.77	5.32
0	0	0	2	1.97	20017	1	1	0	2	7.73	5.33
0	0	0	2	1.82	21801	1	1	0	1	7.66	5.54
0	0	0	2	1.82	21801	1	1	0	2	7.66	5.57
0	0	0	2	1.82	21801	1	1	0	3	7.66	5.57
0	0	0	2	1.82	21801	1	1	0	3	7.66	5.58
0	0	0	2	1.82	21801	1	1	0	4	7.66	5.66
0	0	0	2	1.82	21801	1	1	0	4	7.65	5.67
0	0	0	2	1.82	21801	1	1	0	5	7.65	5.72
0	1	0	1	2.13	27668	2	5	0	1	8.11	4.42
0	1	0	1	2.13	27668	2	5	0	2	8.10	4.37
0	1	0	1	2.13	27668	2	5	0	3	8.08	4.38
0	1	0	4	1.89	18578	2	6	0	1	7.09	5.17
0	1	0	4	1.89	18578	2	6	0	2	7.09	5.24
0	1	0	4	1.89	18578	2	6	0	3	7.56	5.08
1	1	0	4	1.89	18578	2	6	0	3	7.88	5.10
0	0	0	4	2.43	19024	1	6	0	1	7.93	5.15
0	0	0	4	2.43	19024	1	6	0	2	7.92	5.19
0	0	0	4	2.43	19024	1	6	0	3	7.89	5.15
0	0	0	4	2.43	19024	1	6	0	3	7.88	5.11
0	0	0	4	2.43	19025	2	6	0	1	7.93	5.15
0	0	0	4	2.43	19025	2	6	0	2	7.92	5.19

The first few lines and columns of `vwks4_30k.tab`

21.3 Suggested exercise

1. Estimate a cloglog link model (without random effects) for the binary response `match`, treat `t` as a factor variable and include the covariates (`loguu, logvv, nonman, written, size, wage, grade, dayrel`). These results are given as the 1st part (homogeneous model) of the output that is obtained by estimating the random effects model. This is done in Task 2.

2. Allow for a level-2 vacancy random effect (`vacref`), use adaptive quadrature with `mass 48.` Is this random effect significant?

3. Re-estimate the model with the level-2 random effect (`vacref)` and firm (`empref`) as the level 3 random effect. Use adaptive quadrature with `mass 64` for both levels. Are both these random effects significant?

4. How did your results on some important variables e.g. t change, when you allowed for both vacancy-level (level 2) and then firm-level (level 3) random effects?

21.4 References

Andrews, M., Bradley, S., Stott, D., Upward, R., (2007), Testing theories of labour market matching, http://ideas.repec.org/p/ecj/ac2003/209.html.

22 Exercise EP1. Trade Union Membership with Endpoints

The data set we use in this exercise is derived from `nlswork.tab` as described at the start of the Stata, Longitudinal/Panel Data, Release 10, Manual. The data set, `nlswork.tab` is a subsample of the National Longitudinal Survey of Youth data, for the source of the data see http://www.bls.gov/nls/. The Stata subset is for 4711 young women aged 14-26 in 1968, who were then followed for 21 years, excluding the years: 1974, 1976, 1979, 1981, 1984 and 1986. While the Stata datset, `nlswork.tab` had 28534 observations on 21 variables. The `union` variable in this data set only had 19238 non-missing observations. We dropped all observations with missing values on any of the variables used in either the binary response model for `union` or for a linear model of log wage to create our own version of this data. This gave us the dataset, `nls.tab` we use here, it contains 18995 observations on 20 variables (the variables: `ind_code, occ_code, wks_ue, hours` and `wks_work` were dropped from the original dataset as these variables are not used. The variables `black, age2, ttl_exp2` and `tenure2` were created. By dropping specific observations with missing variables rather than dropping all of the observations for each individual with any missing variables, there are more gaps in the `nls.tab` than in `nlswork.tab`. For example, in `nlswork.tab` the individual with idcode 1 is observed in years 1970, 1971, 1972, 1973, 1975, 1977, 1978, 1980, 1983, 1985, 1987 and 1988, whereas in `nls.tab,` this individual is only observed in years 1972, 1977, 1980, 1983, 1985, 1987 and 1988. Gaps do not matter in a repeated cross section models.

22.1 Data description for `nls.tab`

Number of observations: 18995
Number of level-2 cases: 4132

22.2 Variables

`idcode`: NLS id
`year`: interview year
`birth_yr`: birth year
`age`: age in current year
`race`: 1=white, 2=black, 3=other
`msp`: 1 if respondent married and spouse present, 0 otherwise
`nev_mar`: 1 if never yet married, 0 otherwise
`grade`: current grade completed (years of schooling
`collgrad`: 1 if college graduate, 0 otherwise
`not_smsa`: 1 if not SMSA (standard metropolitan statistical area), 0 otherwise
`c_city`: 1 if central city, 0 otherwise
`south`: 1 if South, 0 otherwise
`union`: 1 if union (membership), 0 otherwise
`ttl_exp`: total work experience, 0 otherwise
`tenure`: job tenure, in years
`ln_wage`: ln(wage/GNP deflator)
`black`: 1 if respondent is black, 0 otherwise

`age2`: age squared
`ttl_exp2`: total work experience squared
`tenure2`: tenure squared

idcode	year	birth_yr	age	race	msp	nev_mar	grade	collgrad	not_smsa	c_city	south	union	ttl_exp	tenure	ln_wage	black
1	72	51	20	2	1	0	12	0	0	1	0	1	2.26	0.92	1.59	1
1	77	51	25	2	0	0	12	0	0	1	0	0	3.78	1.50	1.78	1
1	80	51	28	2	0	0	12	0	0	1	0	1	5.29	1.83	2.55	1
1	83	51	31	2	0	0	12	0	0	1	0	1	5.29	0.67	2.42	1
1	85	51	33	2	0	0	12	0	0	1	0	1	7.16	1.92	2.61	1
1	87	51	35	2	0	0	12	0	0	0	0	1	8.99	3.92	2.54	1
1	88	51	37	2	0	0	12	0	0	0	0	1	10.33	5.33	2.46	1
2	71	51	19	2	1	0	12	0	0	1	0	0	0.71	0.25	1.36	1
2	77	51	25	2	1	0	12	0	0	1	0	1	3.21	2.67	1.73	1
2	78	51	26	2	1	0	12	0	0	1	0	1	4.21	3.67	1.69	1
2	80	51	28	2	1	0	12	0	0	1	0	1	6.10	5.58	1.73	1
2	82	51	30	2	1	0	12	0	0	1	0	1	7.67	7.67	1.81	1
2	83	51	31	2	1	0	12	0	0	1	0	1	8.58	8.58	1.86	1
2	85	51	33	2	0	0	12	0	0	1	0	1	10.18	1.83	1.79	1
2	87	51	35	2	0	0	12	0	0	1	0	1	12.18	3.75	1.85	1
2	88	51	37	2	0	0	12	0	0	1	0	1	13.62	5.25	1.86	1
3	71	45	25	2	0	1	12	0	0	1	0	0	3.44	1.42	1.55	1
3	72	45	26	2	0	1	12	0	0	1	0	0	4.44	2.42	1.61	1
3	73	45	27	2	0	1	12	0	0	1	0	0	5.38	3.33	1.60	1
3	77	45	31	2	0	1	12	0	0	1	0	0	6.94	2.42	1.62	1
3	78	45	32	2	0	1	12	0	0	1	0	0	7.98	3.42	1.57	1

First few lines of `nls.tab`

22.3 Suggested exercise

1. Estimate a binary response model for the response variable `union`, with the covariates: `age, age2, black, msp, grade, not_smsa, south, cons`. Use a probit link with adaptive quadrature and mass 36.

2. Reestimate the same model but allow for both lower and upper endpoints. How much of an improvement in log likelihood do you get with the endpoints model? Can the model be simplified? How do you interpret the results of your preferred model?

22.4 References

Stata, Longitudinal/Panel Data, Release 10, Manual (2007), StataCorp, Stata Press, College Station, Texas.

23 Exercise EP2. Poisson Model of the Number of Fish Caught by Visitors to a US National Park.

The data set we use in this exercise is the `fish.tab` as described in the Zero Inflated Poisson Regression Section of the Stata, Reference Q-Z, Release 10, Manual. The data set `fish.tab` contains data on the number of fish caught by parties of visitors to a US National Park, but does not distinguish between parties to the National Park that fish and those that do not. So we might expect that it will include a significant proportion of zero counts made up from those that do not fish and those that did fish but were unsuccessful. In this exercise we will see if a lower endpoint is present in a random effects Poisson model for the number of fish caught.

23.1 Data description for `fish.tab`

Number of observations: 250
Number of level-2 cases: 250

23.2 Variables

`livebait`: 1 if livebait was used, 0 otherwise
`camper`: 1 if the visitors used a camper, 0 otherwsie
`persons`: number of people in the party
`child`: number of children in the party
`count`: number of fish caught
`id`: party identifier

Besides the variables above, the data set `fish.tab` contains covarites that are not used in this analysis.

nofish	livebait	camper	persons	child	xb	zg	count	id
1	0	0	1	0	-0.90	3.05	0	1
0	1	1	1	0	-0.56	1.75	0	2
0	1	0	1	0	-0.40	0.28	0	3
0	1	1	2	1	-0.96	-0.60	0	4
0	1	0	1	0	0.44	0.53	1	5
0	1	1	4	2	1.39	-0.71	0	6
0	1	0	3	1	0.18	-3.40	0	7
0	1	0	4	3	2.33	-5.45	0	8
1	0	1	3	2	0.19	-1.53	0	9
0	1	1	1	0	0.29	1.39	1	10
0	1	0	4	1	1.99	-1.93	0	11
0	1	1	3	2	1.32	-2.47	0	12
1	0	0	3	0	0.30	1.59	1	13
0	1	0	3	0	1.29	0.83	2	14
0	1	1	1	0	-0.06	2.82	0	15
1	1	1	1	0	0.37	2.16	1	16
0	1	0	4	1	1.98	-3.07	0	17
1	1	1	3	2	0.72	-1.95	0	18
0	1	1	2	1	1.52	-0.19	1	19
0	1	0	3	1	-0.03	-0.12	0	20

First few lines and columns of `fish.tab`

23.3 Suggested exercise

1. Estimate a Poisson model for the response variable `count,` with the covariates: `persons, livebait, cons.` Use adaptive quadrature and `mass 36`.

2. Reestimate the same model but allow for lower endpoints. How much of an improvement in loglikelihood do you get with the endpoints model? What happens to your inference on the covariates?

3. How would you interpret the results of your preferred model?

23.4 References

Stata, Reference Q-Z, Release 10, Manual, (2007), StataCorp, Stata Press, College Station, Texas.

24 Exercise EP3. Binary Response Model of Female Employment Participation.

The data set we use in this exercise is from Heckman and Willis (1977). Heckman and Willis (1977) use panel data to investigate the variation in labour force participation rates amongst married women. Their work stemmed from research by Ben-Porath (1973) who observed that cross sectional studies are ambiguous with respect to some important dynamic characteristics of labour force participation. The University of Michigan Panel Study of Income Dynamics 1968-1972 (Morgan et al 1974) provided Heckman and Willis (1977) with employment participation data on white women who were continuously married to the same husband during the 5 year period 1967-1971. A woman was defined as having participated in the labour force in the appropriate year if the respondent answered yes to the question: "Did your wife do any work for money last year". The data, reconstructed from Heckman and Willis (1977) are presented in grouped and long form below: participation in the labour market is coded 1 and non participation is coded 0. This data set in long form (`labour.tab`) was used by Davies, Crouchley and Pickles (1982).

Series	Frequency	Series	Frequency	Series	Frequency	Series	Frequency
0 0 0 0 0	559	1 0 0 1 0	3	1 1 1 0 0	47	0 1 0 1 1	10
1 0 0 0 0	43	1 0 0 0 1	4	1 1 0 1 0	1	0 0 1 1 1	54
0 1 0 0 0	24	0 1 1 0 0	17	1 1 0 0 1	12	1 1 1 1 0	38
0 0 1 0 0	28	0 1 0 1 0	3	1 0 1 1 0	7	1 1 1 0 1	16
0 0 0 1 0	23	0 1 0 0 1	5	1 0 1 0 1	0	1 1 0 1 1	11
0 0 0 0 1	35	0 0 1 1 0	16	1 0 0 1 1	8	1 0 1 1 1	21
1 1 0 0 0	28	0 0 1 0 1	6	0 1 1 1 0	11	0 1 1 1 1	73
1 0 1 0 0	10	0 0 0 1 1	37	0 1 1 0 1	7	1 1 1 1 1	426

Grouped Labour Force participation Data (source: Heckman and Willis, 1977)

24.1 Data description for `labour.tab`

Number of observations: 7915
Number of level-2 cases: 1583

24.2 Variables

`case`: female identifier
`t`: year of the study,
`y`: 1 if employment participation in the year, 0 otherwise

case	t	y
1	1	0
1	2	0
1	3	0
1	4	0
1	5	0
2	1	0
2	2	0
2	3	0
2	4	0
2	5	0
3	1	0
3	2	0
3	3	0
3	4	0
3	5	0
4	1	0
4	2	0
4	3	0
4	4	0
4	5	0
5	1	0

The first few lines of
`labour.tab`

24.3 Suggested exercise

1. Estimate a heterogenous logit model for the response variable `y,` allow for nonstationarity by treating `t` as a factor variable. Use adaptive quadrature with `first.mass=64`.

2. Re-estimate the same model but allow for lower and upper endpoints. How much of an improvement in log likelihood do you get with the endpoints model? How do you interpret your results?

24.4 References

Ben-Porath, Y., (1973), Labour force participation rates and the supply of labour, Journal of Political Economy, 81, 697-704.

Davies, R.B., Crouchley R., and Pickles, A.R., (1982), A family of tests for a collection of short event series with an application to female employment participation, Environment and Planning A, 14, 603-614.

Heckman, J.J., and Willis, R.J., (1977), A beta logistic model for the analysis of sequential labor force participation by married women, Journal of Political

Economy, 85, 27-58.

Morgan, J., Dickinson, K., Dickinson, J., Benus J., Duncam G., (1974), Five Thousand American Families, Patterns of Economic Progress, Volumes 1 and 2, Institute of Social Research, University of Michigan, Ann Arbour, MI.

25 Exercise FOL1. Binary Response Model for Trade Union Membership 1980-1987 of Young Males (Wooldridge, 2005)

Wooldridge (2005) used the data from Vella and Verbeek (1998) on the binary response trade union membership to illustrate his treatment of the initial conditions problem in first order Markov models. We will estimate a range of other models on the same data in this exercise. The Vella and Verbeek (1998) data are from the National Longitudinal Survey (Youth Sample) and consist of a sample of 545 full-time working males who have completed their schooling by 1980 and who are then followed from 1980 to 1987. Trade union membership is determined by the question of whether or not the sampled individual had his wage set in a collective bargaining agreement or not. Wooldridge used the time-constant covariates of `educ` (years of schooling) and `race` (black or not), and the time-varying covariate of marital status.

25.1 Conditional analysis

25.1.1 Data description for `unionjmw1.tab`

Number of observations (rows): 3815:
Number of level-2 cases (`nr`): 545

25.1.2 Variables

`nr`: respondent identifier
`year`: calendar year 1981-1987
`black`: 1 if respondent is classified as black, 0 otherwise
`married`: 1 if respondent is currently married, 0 otherwise
`educ`: years of education
`union`: 1 if wage set by collective bargaining, 0 otherwise in current year
`d81`: 1 if year is 1981, 0 otherwise
`d82`: 1 if year is 1982, 0 otherwise
`d83`: 1 if year is 1983, 0 otherwise
`d84`: 1 if year is 1984, 0 otherwise
`d85`: 1 if year is 1985, 0 otherwise
`d86`: 1 if year is 1986, 0 otherwise
`d87`: 1 if year is 1987, 0 otherwise
`union80`: 1 if wage set by collective bargaining, 0 otherwise in 1980 (initial condition)
`union.1`: lagged 1 year value of union variable
`marravg`: average value of married over 1980-1987
`educu80`: years of education for those in full-time education in 1980
`marr81`: 1 if respondent was married in 1981, 0 otherwise
`marr82`: 1 if respondent was married in 1982, 0 otherwise
`marr83`: 1 if respondent was married in 1983, 0 otherwise
`marr84`: 1 if respondent was married in 1984, 0 otherwise
`marr85`: 1 if respondent was married in 1985, 0 otherwise

`marr86`: 1 if respondent was married in 1986, 0 otherwise
`marr87`: 1 if respondent was married in 1987, 0 otherwise

nr	year	black	married	educ	union	d81	d82	d83	d84	d85	d86	d87	union80	union_1	marravg	educu80	marr81
13	1981	0	0	14	1	1	0	0	0	0	0	0	0	0	0	0	0
13	1982	0	0	14	0	0	1	0	0	0	0	0	0	1	0	0	0
13	1983	0	0	14	0	0	0	1	0	0	0	0	0	0	0	0	0
13	1984	0	0	14	0	0	0	0	1	0	0	0	0	0	0	0	0
13	1985	0	0	14	0	0	0	0	0	1	0	0	0	0	0	0	0
13	1986	0	0	14	0	0	0	0	0	0	1	0	0	0	0	0	0
13	1987	0	0	14	0	0	0	0	0	0	0	1	0	0	0	0	0
17	1981	0	0	13	0	1	0	0	0	0	0	0	0	0	0	0	0
17	1982	0	0	13	0	0	1	0	0	0	0	0	0	0	0	0	0
17	1983	0	0	13	0	0	0	1	0	0	0	0	0	0	0	0	0
17	1984	0	0	13	0	0	0	0	1	0	0	0	0	0	0	0	0
17	1985	0	0	13	0	0	0	0	0	1	0	0	0	0	0	0	0
17	1986	0	0	13	0	0	0	0	0	0	1	0	0	0	0	0	0
17	1987	0	0	13	0	0	0	0	0	0	0	1	0	0	0	0	0
18	1981	0	1	12	0	1	0	0	0	0	0	0	0	0	1	0	1
18	1982	0	1	12	0	0	1	0	0	0	0	0	0	0	1	0	1
18	1983	0	1	12	0	0	0	1	0	0	0	0	0	0	1	0	1
18	1984	0	1	12	0	0	0	0	1	0	0	0	0	0	1	0	1
18	1985	0	1	12	0	0	0	0	0	1	0	0	0	0	1	0	1
18	1986	0	1	12	0	0	0	0	0	0	1	0	0	0	1	0	1
18	1987	0	1	12	0	0	0	0	0	0	0	1	0	0	1	0	1
45	1981	0	0	12	1	1	0	0	0	0	0	0	1	1	0.125	12	0
45	1982	0	0	12	0	0	1	0	0	0	0	0	1	1	0.125	12	0
45	1983	0	0	12	0	0	0	1	0	0	0	0	1	0	0.125	12	0

First few lines of `unionjmw1.tab`

25.1.3 Suggested exercise

1. Estimate a random effect probit model (adaptive quadraure, mass 24) of trade union membership (`union`), with a constant, the lagged union membership variable (`union_1`), `educ`, `black` and the marital status dummy variable (`married`), the marr81-marr87 and the d82-d87 sets of dummy variables.

2. Add the initial condition of trade union membership in 1980 (`union80`) to the previous model. How does the inference on the lagged responses (`union_1`) and the scale parameters differ between the two models?

25.2 Joint analysis of the initial condition and subsequent responses

25.2.1 Data description for `unionjmw2.tab`

Number of observations (rows): 4360
Number of level-2 cases (`nr`): 545

25.2.2 Variables

The variables are the same as `unionjmw2.tab` with the addition of `d`, `d1` and `d2` at the end of the list, where:

`d`: 1 for the initial response, 2 if a subsequent response
`d1`: 1 if `d`=1, 0 otherwise
`d2`: 1 if `d`=2, 0 otherwise

nr	year	black	married	educ	union	d81	d82	d83	d84	d85	d86	d87	union80	union_1	marravg	educu80	marr81
13	1980	0	0	14	0	0	0	0	0	0	0	0	-9	-9	0	-9	-9
13	1981	0	0	14	1	1	0	0	0	0	0	0	0	0	0	0	0
13	1982	0	0	14	0	0	1	0	0	0	0	0	0	1	0	0	0
13	1983	0	0	14	0	0	0	1	0	0	0	0	0	0	0	0	0
13	1984	0	0	14	0	0	0	0	1	0	0	0	0	0	0	0	0
13	1985	0	0	14	0	0	0	0	0	1	0	0	0	0	0	0	0
13	1986	0	0	14	0	0	0	0	0	0	1	0	0	0	0	0	0
13	1987	0	0	14	0	0	0	0	0	0	0	1	0	0	0	0	0
17	1980	0	0	13	0	0	0	0	0	0	0	0	-9	-9	0	-9	-9
17	1981	0	0	13	0	1	0	0	0	0	0	0	0	0	0	0	0
17	1982	0	0	13	0	0	1	0	0	0	0	0	0	0	0	0	0
17	1983	0	0	13	0	0	0	1	0	0	0	0	0	0	0	0	0
17	1984	0	0	13	0	0	0	0	1	0	0	0	0	0	0	0	0
17	1985	0	0	13	0	0	0	0	0	1	0	0	0	0	0	0	0
17	1986	0	0	13	0	0	0	0	0	0	1	0	0	0	0	0	0
17	1987	0	0	13	0	0	0	0	0	0	0	1	0	0	0	0	0
18	1980	0	1	12	0	0	0	0	0	0	0	0	-9	-9	1	-9	-9
18	1981	0	1	12	0	1	0	0	0	0	0	0	0	0	1	0	1
18	1982	0	1	12	0	0	1	0	0	0	0	0	0	0	1	0	1
18	1983	0	1	12	0	0	0	1	0	0	0	0	0	0	1	0	1
18	1984	0	1	12	0	0	0	0	1	0	0	0	0	0	1	0	1
18	1985	0	1	12	0	0	0	0	0	1	0	0	0	0	1	0	1
18	1986	0	1	12	0	0	0	0	0	0	1	0	0	0	1	0	1
18	1987	0	1	12	0	0	0	0	0	0	0	1	0	0	1	0	1
45	1980	0	0	12	1	0	0	0	0	0	0	0	-9	-9	0.125	-9	-9
45	1981	0	0	12	1	1	0	0	0	0	0	0	1	1	0.125	12	0

First few lines of `unionjmw2.tab`

25.2.3 Suggested exercise

3 Estimate a common random effect common scale parameter joint probit model (adaptive quadrature, mass 24) of trade union membership (`union_1`). Use the `d1` and `d2` dummy variables to set up the linear predictors. Use constants in both linear predictors. For the initial response, use the `married`, `educ` and `black` regressors. For the subsequent response, use the regressors: lagged union membership variable (`union_1`), `educ`, `black` and the marital status dummy variable (`married`), the `marr81`-`marr87` and the `year` dummy variables. What does this model suggest about state dependence and unobserved heterogeneity?

4 Re-estimate the model allowing the scale parameters for the initial and subsequent responses to be different. Is this a significant improvement over the common scale parameter model?

5 To the different scale parameter model, add the baseline response (`union80`). Does this make a significant improvement to the model?

25.3 References

Vella, F., Verbeek, M., (1998), Whose wages do Unions raise? A dynamic Model of Unionism and wage rate determination for young men, Journal of Applied Econometrics, 13, 163-183.

Wooldridge, J.M., (2005), Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity, Journal of Applied Econometrics, 20, 39-54.

26 Exercise FOL2. Probit Model for Trade Union Membership of Females

This exercise uses a form of the data from the union data for US young women from the National Longitudinal Survey of Youth (NLSY) of the Stata manual (`http://www.stata-press.com/data/r9/union.dta`). We use the same subsample that was used by Stewart (2006) to illustrate his Stata program (redprob). To form this subsample Stewart (2006) uses only data from 1978 onwards; the data for 1983 are dropped, and only those individuals observed in each of the remaining 6 waves are kept. This gave a balanced panel with N = 799 individuals observed in each of I = 6 waves. The observations for 1985 and 1987 are implicitly treated as if they were for 1984 and 1986 respectively, which would give 6 waves at regular 2-year intervals. Trade union membership is determined by the question of whether of not the sampled individual had her wage set in a collective bargaining agreement or not.

26.1 Conditional analysis

26.1.1 Data description for `unionred1.tab`

Number of observations: 3995
Number of level-2 cases: 799

26.1.2 Variables

`idcode`: NLSY subject identifier code
`year`: interview year
`age`: age in current year
`grade`: years of schooling completed
`not.smsa`: 1 if living outside a standard metropolitan statistical area, 0 otherwise
`south`: 1 if south, 0 otherwise
`union`: 1 if wage is collectively negotiated, 0 otherwise
`t0`: `year`-70
`southxt`: 1 if resident in south, 0 otherwise
`black`: 1 if respondent's race black, 0 otherwise
`tper`: panel wave
`lagunion`: the value of `union` in the previous interval
`d`: 2 for all responses, as all responses are post baseline.
`d1`: 0 for all responses, as all responses are post baseline
`d2`: 1 for all responses, as all responses are post baseline
`baseunion`: 1 if `union`=1 in 1978, 0 otherwise

idcode	year	age	grade	not_smsa	south	union	t0	southXt	black	tper	lagunion	d	d1	d2	baseunion
2	80	28	12	0	0	1	10	0	1	2	1	2	0	1	1
2	82	30	12	0	0	1	12	0	1	3	1	2	0	1	1
2	85	33	12	0	0	1	15	0	1	4	1	2	0	1	1
2	87	35	12	0	0	1	17	0	1	5	1	2	0	1	1
2	88	37	12	0	0	1	18	0	1	6	1	2	0	1	1
3	80	34	12	0	0	0	10	0	1	2	0	2	0	1	0
3	82	36	12	0	0	0	12	0	1	3	0	2	0	1	0
3	85	39	12	0	0	0	15	0	1	4	0	2	0	1	0
3	87	41	12	0	0	0	17	0	1	5	0	2	0	1	0
3	88	42	12	0	0	0	18	0	1	6	0	2	0	1	0
6	80	33	12	0	0	0	10	0	0	2	1	2	0	1	1
6	82	35	12	0	0	0	12	0	0	3	0	2	0	1	1
6	85	38	12	0	0	0	15	0	0	4	0	2	0	1	1
6	87	40	12	0	0	0	17	0	0	5	0	2	0	1	1
6	88	42	12	0	0	0	18	0	0	6	0	2	0	1	1
9	80	28	12	0	0	1	10	0	0	2	1	2	0	1	1
9	82	30	12	0	0	1	12	0	0	3	1	2	0	1	1
9	85	33	12	0	0	1	15	0	0	4	1	2	0	1	1
9	87	35	12	0	0	1	17	0	0	5	1	2	0	1	1
9	88	37	12	0	0	1	18	0	0	6	1	2	0	1	1
13	80	32	14	0	0	1	10	0	0	2	0	2	0	1	0
13	82	34	14	0	0	0	12	0	0	3	1	2	0	1	0
13	85	37	14	0	0	0	15	0	0	4	0	2	0	1	0
13	87	39	14	0	0	0	17	0	0	5	0	2	0	1	0
13	88	40	14	0	0	0	18	0	0	6	0	2	0	1	0
15	80	31	16	0	0	1	10	0	0	2	0	2	0	1	0
15	82	33	16	0	0	1	12	0	0	3	1	2	0	1	0
15	85	36	16	0	0	1	15	0	0	4	1	2	0	1	0
15	87	38	16	0	0	1	17	0	0	5	1	2	0	1	0
15	88	39	16	0	0	1	18	0	0	6	1	2	0	1	0

First few lines of `unionred1.tab`

26.1.3 Suggested exercise

1. Estimate a heterogenous probit (level-2 with idcode, adaptive quadrature, mass 16) model of trade union membership (`union`), with a constant and the lagged union membership variable (`lagunion`), `age`, `grade`, and `southxt` regressors.

2. Add the initial condition of trade union membership in 1978 (`baseunion`) to the previous model. How do the inference on the lagged responses (`lagunion`) and the scale effects differ between the two models.

26.2 Joint analysis of the initial condition and subsequent responses

26.2.1 Data description for `unionred2.tab`

Number of observations: 4794
Number of level-2 cases: 799

26.2.2 Variables

The variables are the same as `unionred2.tab` except that this time the variables `d`, `d1` and `d2` take more values.

`d`: 1, for the initial response, 2 if a subsequent response
`d1`: 1 if `d`=1, 0 otherwise
`d2`: 1 if `d`=2, 0 otherwise

idcode	year	age	grade	not_smsa	south	union	t0	southXt	black	tper	lagunion	d	d1	d2	baseunion
2	78	26	12	0	0	1	8	0	1	1	-9	1	1	0	1
2	80	28	12	0	0	1	10	0	1	2	1	2	0	1	1
2	82	30	12	0	0	1	12	0	1	3	1	2	0	1	1
2	85	33	12	0	0	1	15	0	1	4	1	2	0	1	1
2	87	35	12	0	0	1	17	0	1	5	1	2	0	1	1
2	88	37	12	0	0	1	18	0	1	6	1	2	0	1	1
3	78	32	12	0	0	0	8	0	1	1	-9	1	1	0	0
3	80	34	12	0	0	0	10	0	1	2	0	2	0	1	0
3	82	36	12	0	0	0	12	0	1	3	0	2	0	1	0
3	85	39	12	0	0	0	15	0	1	4	0	2	0	1	0
3	87	41	12	0	0	0	17	0	1	5	0	2	0	1	0
3	88	42	12	0	0	0	18	0	1	6	0	2	0	1	0
6	78	31	12	0	0	1	8	0	0	1	-9	1	1	0	1
6	80	33	12	0	0	0	10	0	0	2	1	2	0	1	1
6	82	35	12	0	0	0	12	0	0	3	0	2	0	1	1
6	85	38	12	0	0	0	15	0	0	4	0	2	0	1	1
6	87	40	12	0	0	0	17	0	0	5	0	2	0	1	1
6	88	42	12	0	0	0	18	0	0	6	0	2	0	1	1
9	78	26	12	0	0	1	8	0	0	1	-9	1	1	0	1
9	80	28	12	0	0	1	10	0	0	2	1	2	0	1	1
9	82	30	12	0	0	1	12	0	0	3	1	2	0	1	1
9	85	33	12	0	0	1	15	0	0	4	1	2	0	1	1
9	87	35	12	0	0	1	17	0	0	5	1	2	0	1	1
9	88	37	12	0	0	1	18	0	0	6	1	2	0	1	1
13	78	30	14	0	0	0	8	0	0	1	-9	1	1	0	0
13	80	32	14	0	0	1	10	0	0	2	0	2	0	1	0
13	82	34	14	0	0	0	12	0	0	3	1	2	0	1	0
13	85	37	14	0	0	0	15	0	0	4	0	2	0	1	0
13	87	39	14	0	0	0	17	0	0	5	0	2	0	1	0

First few lines of `unionred2.tab`

26.2.3 Suggested exercise

3 Estimate a common random effect common scale joint probit model (use adapptive quadrature `mass 24`) of trade union membership (`union`). Use constants in both linear predictors. Use the `d1` and `d2` dummy variables to set up the linear predictors. For the initial response use the regressors: `age`, `grade`, `southxt` and `not_smsa`. For the subsequent response use the regressors: lagged union membership variable (`lagunion`), `age`, `grade`, `southxt`. What does this model suggest about state dependence and unobserved heterogeneity?

4 Re-estimate the model allowing the scale parameters for the initial and subsequent responses to be different (use adaptive quadrature with `mass 32`). Is this a significant improvement over the common scale parameter model?

5 Re-estimate the model using a bivariate model for the random effects (common scale). Are these results different to those of Task 4?

6 To the bivariate model of Task 5 add the initial or baseline response (`baseunion`). Are these results different to those of Task 5?

26.3 References

Stewart, M.B., (2006), -redprob- A Stata program for the Heckman estimator of the random effects dynamic probit model,
http://www2.warwick.ac.uk/fac/soc/economics/staff/faculty/stewart/stata/redprobnote.pdf.

27 Exercise FOL3. Binary Response Model for Female Labour Force Participation in the UK

Davies, Elias and Penn (1992) and Davies (1993) as part of the ESRC funded Social Change and Economic Life Initiative. The data we use is the annual employment behaviour of wives from Rochdale (UK) from the date of their marriage to the end of the survey in 1987. The binary response `femp` takes the value 1 if a wife was employed in the current year and 0 otherwise. There is a set of explanatory variables that include husband's employment status and age (years). In this exercise we are going to see if we can distinguish state dependence (1^{st} order effects) in employment behaviour of wives from unobserved heterogeneity. Versions of the same data (`wemp.tab`) were used by Rabe-Hesketh and Skrondal (2005, exercise 4.5).

27.1 Conditional analysis

27.1.1 Data description for `wemp-base1.tab`

Number of observations: 1274
Number of level-2 cases: 144

27.1.2 Variables

`case`: identifier for wives
`femp`: 1 if wife is in employment status in current year, 0 otherwise
`mune`: 1 if the husband is in employment in current year, 0 otherwise
`time`: year of observation-1975
`und1`: 1 if the wife has children under the age of 1, 0 otherwise
`und5`: 1 if the wife has children under the age of 5, 0 otherwise
`age`: wife's age-1975
`ylag`: femp lagged 1 year
`ybase`: femp in 1st year
`r`: 2 for allpost 1st year observations
`r1`: 0 for all observations
`r2`: 1, if r=2

The data set contains variables not used in this analysis.

case	femp	mune	time	und1	und5	age	d	d1	d0	ylag	ybase	r	r1	r2
1	0	0	11	0	1	-10	1	0	1	0	0	2	0	1
1	0	0	12	0	1	-9	1	0	1	0	0	2	0	1
6	0	0	1	0	0	9	2	1	0	1	1	2	0	1
6	0	0	2	0	0	10	1	0	1	0	1	2	0	1
6	1	0	3	0	0	11	1	0	1	0	1	2	0	1
6	1	0	4	0	0	12	2	1	0	1	1	2	0	1
6	1	0	5	0	0	13	2	1	0	1	1	2	0	1
6	1	0	6	0	0	14	2	1	0	1	1	2	0	1
6	1	0	7	0	0	15	2	1	0	1	1	2	0	1
6	1	0	8	0	0	16	2	1	0	1	1	2	0	1
6	1	0	9	0	0	17	2	1	0	1	1	2	0	1
6	1	0	10	0	0	18	2	1	0	1	1	2	0	1
6	1	0	11	0	0	19	2	1	0	1	1	2	0	1
6	1	0	12	0	0	20	2	1	0	1	1	2	0	1
20	1	0	7	0	0	-9	2	1	0	1	1	2	0	1
20	0	0	8	1	1	-8	2	1	0	1	1	2	0	1
20	1	0	9	0	1	-7	1	0	1	0	1	2	0	1
20	1	0	10	0	1	-6	2	1	0	1	1	2	0	1
20	1	0	11	0	1	-5	2	1	0	1	1	2	0	1
20	1	0	12	1	1	-4	2	1	0	1	1	2	0	1
24	1	0	1	0	0	0	2	1	0	1	1	2	0	1
24	1	0	2	0	0	1	2	1	0	1	1	2	0	1
24	1	0	3	0	0	2	2	1	0	1	1	2	0	1
24	1	0	4	0	0	3	2	1	0	1	1	2	0	1
24	1	0	5	0	0	4	2	1	0	1	1	2	0	1
24	1	0	6	0	0	5	2	1	0	1	1	2	0	1

First few lines of `wemp-base1.tab`

27.1.3 Suggested exercise

1. Estimate a heterogenous logit (level-2 with `case`, use adaptive quadrature, `mass 12`) model of female employment participation (`femp`), with a constant and the lagged female employment participation variable (`ylag`), `mune`, `und5`, and `age` regressors.

2. Add the initial condition of employed in the 1st year (`ybase`) to the previous model. How do the inference on the lagged responses (`ylag`) and the scale effects differ between the two models.

27.2 Joint analysis of the initial condition and subsequent responses

27.2.1 Data description for `wemp-base2.tab`

Number of observations: 1425
Number of level-2 cases: 151

27.2.2 Variables

The variables are the same as `wemp-base2.tab` except that this time the variables `ylag`, `r`, `r1` and `r2` take more values

`ylag`: femp lagged 1 year, -9 if its the 1st year
`r`: 1 for the initial response, 2 if a subsequent response
`r1`: 1 if d=1, 0 otherwise
`r2`: 1 if d=2, 0 otherwise

case	femp	mune	time	und1	und5	age	d	d1	d0	ylag	ybase	r	r1	r2
1	0	0	10	0	1	-11	2	1	0	-9	0	1	1	0
1	0	0	11	0	1	-10	1	0	1	0	0	2	0	1
1	0	0	12	0	1	-9	1	0	1	0	0	2	0	1
6	1	0	0	0	0	8	2	1	0	-9	1	1	1	0
6	0	0	1	0	0	9	2	1	0	1	1	2	0	1
6	0	0	2	0	0	10	1	0	1	0	1	2	0	1
6	1	0	3	0	0	11	1	0	1	0	1	2	0	1
6	1	0	4	0	0	12	2	1	0	1	1	2	0	1
6	1	0	5	0	0	13	2	1	0	1	1	2	0	1
6	1	0	6	0	0	14	2	1	0	1	1	2	0	1
6	1	0	7	0	0	15	2	1	0	1	1	2	0	1
6	1	0	8	0	0	16	2	1	0	1	1	2	0	1
6	1	0	9	0	0	17	2	1	0	1	1	2	0	1
6	1	0	10	0	0	18	2	1	0	1	1	2	0	1
6	1	0	11	0	0	19	2	1	0	1	1	2	0	1
6	1	0	12	0	0	20	2	1	0	1	1	2	0	1
20	1	0	6	0	0	-10	2	1	0	-9	1	1	1	0
20	1	0	7	0	0	-9	2	1	0	1	1	2	0	1
20	0	0	8	1	1	-8	2	1	0	1	1	2	0	1
20	1	0	9	0	1	-7	1	0	1	0	1	2	0	1
20	1	0	10	0	1	-6	2	1	0	1	1	2	0	1
20	1	0	11	0	1	-5	2	1	0	1	1	2	0	1
20	1	0	12	1	1	-4	2	1	0	1	1	2	0	1
24	1	0	0	0	0	-1	2	1	0	-9	1	1	1	0
24	1	0	1	0	0	0	2	1	0	1	1	2	0	1
24	1	0	2	0	0	1	2	1	0	1	1	2	0	1
24	1	0	3	0	0	2	2	1	0	1	1	2	0	1

First few lines of `wemp-base2.tab`

27.2.3 Suggested exercise

3 Estimate a common random effect common scale joint logit model (adaptive quadrature, `mass 12`) of female employment participation (`femp`). Use constants in both linear predictors. Use the `r1` and `r2` dummy variables to set up the linear predictors. For the initial response use the regressors: `mune`, `und5`, and `age` regressors. For the subsequent responses use the regressors: the lagged female employment participation variable (`ylag`), `mune`, `und5`, and `age`. What does this model suggest about state dependence and unobserved heterogeneity?

4 Re-estimate the model allowing the scale parameters for the initial and subsequent responses to be different.

5 In this model, replace the lagged female employment participation variable (`ylag`) with the initial or baseline response (`ybase`). Are these results different to those of Task 4?

6 I In this model, include both the lagged response (`ylag`) and the baseline response (`ybase`). Are these results different to those of Task 5?

7 Re-estimate the model with the baseline response (`ybase`) and the lagged response (`ylag`) using a bivariate model for the random effects (common scale).

8 Compare the results obtained for the various models on the covariates and role of employment status in the previous year. Are both state dependence and unobserved heterogeneity present in this data?

27.3 References

Davies, R.B., Elias, P., and Penn, R., (1992), The relationship between a husband's unemployment and his wife's participation in the labour force, Oxford Bulletin of Economics and Statistics, 54, 145-171.

Davies, R.B., (1993), Statistical modelling for survey analysis, Journal of the Market Research Society, 35, 235-247.

Rabe-Hesketh, S., & Skrondal, A., (2005), Multilevel and Longitudinal Modelling using Stata, Stata Press, Stata Corp, College Station, Texas.

28 Exercise FOC4. Poisson Model of Patents and R&D Expenditure

The data we use in this example are from Hall, Griliches Hausman (1986), the data refer to the number of Patents awarded to a sample of 346 firms each year from 1975 to 1979. Hall et al (1986) were particularly interested in the effect of current and lagged research and development (R&D) expenditures on the number of awarded patents. The data we use here (`patents.tab`) are a version of that made available by Cameron and Trivedi (1988). All spending in the data set is in 1972 US dollars.

28.1 Data description for `patents.tab`

Number of observations: 1680
Number of level-2 cases: 336, the original data was for 346 firms

28.2 Variables

`obsno`: firm identifier (1,2,...,336)
`year`: year identifier, 1=1975, 2=1976, 3=1977, 4=1978, 5=1979
`cusip`: Compustat's identifying number for the firm
`ardssic`: a two-digit code for the applied R&D industrial classification
`scisect`: 1 for firms in the scientific sector, 0 otherwise
`logk`: the logarithm of the book value of the firms's capital value in 1972.
`sumpat`: the sum of patents applied for between 1972-1979.
`pat`: the number of patents applied for during the current year that were eventually granted.
`pat1`: the number of patents applied for during the previous year that were eventually granted.
`pat2`: the number of patents applied for two years ago that were eventually granted.
`pat3`: the number of patents applied for three years ago that were eventually granted.
`pat4`: the number of patents applied for four years ago that were eventually granted.
`logr`: the logarithm of R&D spending
`logr1`: the logarithm of R&D spending in previous year
`logr2`: the logarithm of R&D spending 2 years ago
`logr3`: the logarithm of R&D spending 3 years ago
`logr4`: the logarithm of R&D spending 4 years ago
`logr5`: the logarithm of R&D spending 5 yeras ago
`year1`: 1 for year=1975, 0 otherwise
`year2`: 1 for year=1976, 0 otherwise
`year3`: 1 for year=1977, 0 otherwise
`year4`: 1 for year=1978, 0 otherwise

`year5`: 1 for year=1979, 0 otherwise
`r`: 1 if the the current year is the base-line year, 2 othervise
`r1`: 1 if r=1, 0 otherwise
`r2`: 1 if r=2, 0 otherwise

obsno	year	cusip	ardssic	scisect	logk	sumpat	pat	pat1	pat2	pat3	pat4	logr	logr1	logr2	logr3	logr4	logr5	year1	year2	year3	year4	year5	r	r1	r2	base
1	1	800	15	0	6.08	354	32	31	34	22	28	0.92	1.03	1.07	0.94	0.88	1.00	1	0	0	0	0	1	1	0	32
1	2	800	15	0	6.08	354	41	32	31	34	22	1.02	0.92	1.03	1.07	0.94	0.88	0	1	0	0	0	2	0	1	32
1	3	800	15	0	6.08	354	60	41	32	31	34	0.97	1.02	0.92	1.03	1.07	0.94	0	0	1	0	0	2	0	1	32
1	4	800	15	0	6.08	354	57	60	41	32	31	1.10	0.97	1.02	0.92	1.03	1.07	0	0	0	1	0	2	0	1	32
1	5	800	15	0	6.08	354	77	57	60	41	32	1.08	1.10	0.97	1.02	0.92	1.03	0	0	0	0	1	2	0	1	32
2	1	1030	14	1	1.97	13	3	2	1	2	1	-1.49	-0.68	-0.15	0.08	-0.22	-0.46	1	0	0	0	0	1	1	0	3
2	2	1030	14	1	1.97	13	2	3	2	1	2	-1.19	-1.49	-0.68	-0.15	0.08	-0.22	0	1	0	0	0	2	0	1	3
2	3	1030	14	1	1.97	13	1	2	3	2	1	-0.61	-1.19	-1.49	-0.68	-0.15	0.08	0	0	1	0	0	2	0	1	3
2	4	1030	14	1	1.97	13	1	1	2	3	2	-0.58	-0.61	-1.19	-1.49	-0.68	-0.15	0	0	0	1	0	2	0	1	3
2	5	1030	14	1	1.97	13	1	1	1	2	3	-0.61	-0.58	-0.61	-1.19	-1.49	-0.68	0	0	0	0	1	2	0	1	3
3	1	2824	4	1	5.65	493	49	58	63	61	43	3.67	3.59	3.53	3.44	3.41	3.39	1	0	0	0	0	1	1	0	49
3	2	2824	4	1	5.65	493	42	49	58	63	61	3.78	3.67	3.59	3.53	3.44	3.41	0	1	0	0	0	2	0	1	49
3	3	2824	4	1	5.65	493	63	42	49	58	63	3.82	3.78	3.67	3.59	3.53	3.44	0	0	1	0	0	2	0	1	49
3	4	2824	4	1	5.65	493	77	63	42	49	58	3.88	3.82	3.78	3.67	3.59	3.53	0	0	0	1	0	2	0	1	49
3	5	2824	4	1	5.65	493	80	77	63	42	49	3.91	3.88	3.82	3.78	3.67	3.59	0	0	0	0	1	2	0	1	49
4	1	4644	13	0	0.68	2	0	0	1	0	0	0.43	0.54	0.49	0.59	0.48	0.54	1	0	0	0	0	1	1	0	0
4	2	4644	13	0	0.68	2	0	0	0	1	0	0.34	0.43	0.54	0.49	0.59	0.48	0	1	0	0	0	2	0	1	0

The first few lines of `patents.tab`

28.3 Suggested exercise

1. We are going to estimate several versions of the joint model of the initial and subsequent responses, to do this we will want the covariates to have different parameter estimates in the model for the initial conditions to those we want to obtain for the subsequent responses. This implies that we will need to create interaction effects with the `r1` and `r2` indicators.

2. The 1st model to be estimated has a common random effect for the baseline and subsequent responses but excludes the lagged response. Use the covariates: `r1, r1_logr, r1_logk, r1_scisect` for the baseline, and the covariates `r2, r2_logr, r2_logk, r2_scisect, r2_year3, r2_year4, r2_year5` for the subsequent responses. Use adaptive quadrature and `first.mass=36`. Add the previous outcome, `r2_pat1` to establish if we have a 1st order model. If this is significant we can add `r2_base` to establish whether the Wooldridge (2005) control adds anything to the model. Interpret your results?

3. Repeat Task 2 with a 1 factor model for the baseline and subsequent responses with adaptive quadrature, `first.mass=24` and accurate arithmetic.

4. Repeat Task 3 using a bivariate model for the baseline and subsequent responses with adaptive quadrature, `first.mass=36` in both dimensions and with accurate arithmetic.

5. Compare the results, which is your preferred model and why?

28.4 References

Hall, B., Griliches, Z., and Hausman, J., (1986), Patents and R&D: Is There a Lag?, International Economic Review, 27, 265-283.

Cameron, A.C., and Trivedi, P.K., (1998), Regression Analysis of Count Data, Econometric Society Monograph No.30, Cambridge University Press, see http://cameron.econ.ucdavis.edu/racd/racddata.html.

Wooldridge, J.M., (2005), Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity, Journal of Applied Econometrics, 20, 39—54.

29 Exercise FE1. Linear Model for the Effect of Job Training on Firm Scrap Rates

Holzer, Block, Cheatham and Knott (1993) studied the impact of job training grants on worker productivity by collecting information on "scrap rates" for a sample of Michigan manufacturing firms. In a related study Wooldridge (2006, Example 14.1) uses data (`jtrain.tab`) on 54 firms that reported "scrap rates" for the years 1987, 1988 and 1989. No firms obtained job training grants before 1988, 19 firms obtained grants in 1989. Wooldridge (2006) allowed for the possibility that the additional job training in 1988 made workers more productive in 1989 by use of the lagged value of the grant indicator, he also included indicator variables for the 1988 and 1989. We will replicate the Wooldridge (2006) analysis in this exercise.

29.1 Data description for `jtrain.tab`

Number of observations: 162
Number of level-2 cases: 54

29.2 Variables

`year`: 1987, 1988, or 1989
`fcode`: firm code number
`employ`: number of employees at plant
`sales`: annual sales, $
`avgsal`: average employee salary
`scrap`: scrap rate (per 100 items)
`rework`: rework rate (per 100 items)
`tothrs`: total hours training
`union`: 1 if firm unionized, 0 otherwise
`grant` 1 if firm received grant, 0 otherwise
`d89`: 1 if year = 1989, 0 otherwise
`d88`: 1 if year = 1988, 0 otherwise
`totrain`: total employees trained
`hrsemp`: tothrs/totrain
`lscrap`: log(scrap)
`lemploy`: log(employ)
`lsales`: log(sales)
`lrework`: log(rework)
`lhrsemp`: log(1 + hrsemp)
`lscrap_1`: lagged lscrap; missing 1987
`grant_1`: lagged grant; assumed 0 in 1987
`clscrap`: lscrap - lscrap_1; year > 1987
`cgrant`: grant - grant_1
`clemploy`: lemploy - lemploy[t-1]

`clsales`: lavgsal - lavgsal[t-1]
`lavgsal`: log(avgsal)
`clavgsal`: lavgsal - lavgsal[t-1]
`cgrant_1`: cgrant[t-1]
`chrsemp`: hrsemp - hrsemp[t-1]
`clhrsemp`: lhrsemp - lhrsemp[t-1]

year	fcode	employ	sales	avgsal	scrap	rework	tothrs	union	grant	d89	d88	totrain	hrsemp	lscrap
1987	410032	100	47000000	35000			12	0	0	0	0	100	12.00	
1988	410032	131	43000000	37000			8	0	0	0	1	50	3.05	
1989	410032	123	49000000	39000			8	0	0	1	0	50	3.25	
1987	410440	12	1560000	10500			12	0	0	0	0	12	12.00	
1988	410440	13	1970000	11000			12	0	0	0	1	13	12.00	
1989	410440	14	2350000	11500			10	0	0	1	0	14	10.00	
1987	410495	20	750000	17680			50	0	0	0	0	15	37.50	
1988	410495	25	110000	18720			50	0	0	0	1	10	20.00	
1989	410495	24	950000	19760			50	0	0	1	0	20	41.67	
1987	410500	200	23700000	13729			0	0	0	0	0	0	0.00	
1988	410500	155	19700000	14287			0	0	0	0	1	0	0.00	
1989	410500	80	26000000	15758			24	0	0	1	0	20	6.00	
1987	410501		6000000				0	0	0	0	0	10		
1988	410501		8000000				0	0	0	0	1	20		
1989	410501		10000000				0	0	0	1	0	25		
1987	410509						0	0	0	0	0	0		
1988	410509		2800000	18000			0	0	0	0	1	0		
1989	410509	20	3400000	18500			0	0	0	1	0	0	0.00	

First few lines and columns of `jtrain.tab`

29.3 Suggested exercise

1. Estimate a homogeneous linear model for the response `lscrap`, with covariates `grant, d89, d88 and grant_1`. These results are given as the 1st part (homogeneous model) of the output that is obtained by estimating the fixed effects model. Estimate the model using the fixed firm effects (`fcode`). What is the main difference between the results from the alternative estimators?

2. Re-estimate the models of Task 1 without the lagged grant indicator (`grant_1`). Is the model a poorer fit to the data?

3. What does the coefficient for `d89` suggest in your preferred model?

4. Re-estimate the fixed effects models of questions 1 and 2 using adaptive quadrature and `first.mass= 64`. Compare the fixed and random effect model inferences. What do you find?

29.4 References

Holzer, H., Block, R., Cheatham, M., and Knott, J., (1993), Are training subsidies effective? The Michigan experience, Industrial and Labor Relations Review, 46, 625-636.

Wooldridge, J. M. (2006), Introductory Econometrics: A Modern Approach. Third edition. Thompson, Australia.

30 Exercise FE2. Linear Model to Establish if the Returns to Education Changed over Time

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. Some of the variables change over time, three important ones are: years of labour market experience, marital status, and trade union membership. On the other hand some variables such as: race, education do not change. Following Wooldridge (2006, Example 14.44) we use a version of the Vella and Verbeek (1998) data (`wagepan2.tab`), in various models of the response variable, log wages.

30.1 Data description for `wagepan2.tab`

Number of observations: 4360
Number of level-2 cases: 545

30.2 Variables

`nr`: person identifier
`year`: 1980 to 1987
`black`: 1 if respondent is black, 0 otherwise
`exper`: labor mkt experience
`hisp`: 1 if respondent is Hispanic, 0 otherwise
`hours`: annual hours worked
`married`: 1 if respondent is married, 0 otherwise
`educ`: years of schooling
`union`: 1 if respondent is in union, 0 otherwise
`lwage`: log(wage)
`d81`: 1 if year = 1981, 0 otherwise
`d82`: 1 if year = 1982, 0 otherwise
`d83`: 1 if year = 1983, 0 otherwise
`d84`: 1 if year = 1984, 0 otherwise
`d85`: 1 if year = 1985, 0 otherwise
`d86`: 1 if year = 1986, 0 otherwise
`d87`: 1 if year = 1987, 0 otherwise
`expersq`: exper^2

The data set (`wagepan2.tab)` includes other variables that are not used in this analysis.

nr	year	black	exper	hisp	hours	married	occ1	occ2	occ3	occ4	occ5	occ6	occ7	occ8	occ9	educ	union	lwage	d81	d82	d83	d84	d85	d86	d87	expersq
13	1980	0	1	0	2672	0	0	0	0	0	0	0	0	0	1	14	0	1.20	0	0	0	0	0	0	0	1
13	1981	0	2	0	2320	0	0	0	0	0	0	0	0	0	1	14	1	1.85	1	0	0	0	0	0	0	4
13	1982	0	3	0	2940	0	0	0	0	0	0	0	0	0	1	14	0	1.34	0	1	0	0	0	0	0	9
13	1983	0	4	0	2960	0	0	0	0	0	0	0	0	0	1	14	0	1.43	0	0	1	0	0	0	0	16
13	1984	0	5	0	3071	0	0	0	0	0	1	0	0	0	0	14	0	1.57	0	0	0	1	0	0	0	25
13	1985	0	6	0	2864	0	0	1	0	0	0	0	0	0	0	14	0	1.70	0	0	0	0	1	0	0	36
13	1986	0	7	0	2994	0	0	1	0	0	0	0	0	0	0	14	0	-0.72	0	0	0	0	0	1	0	49
13	1987	0	8	0	2640	0	0	1	0	0	0	0	0	0	0	14	0	1.67	0	0	0	0	0	0	1	64
17	1980	0	4	0	2484	0	0	1	0	0	0	0	0	0	0	13	0	1.68	0	0	0	0	0	0	0	16
17	1981	0	5	0	2804	0	0	1	0	0	0	0	0	0	0	13	0	1.52	1	0	0	0	0	0	0	25
17	1982	0	6	0	2530	0	0	1	0	0	0	0	0	0	0	13	0	1.56	0	1	0	0	0	0	0	36
17	1983	0	7	0	2340	0	0	1	0	0	0	0	0	0	0	13	0	1.73	0	0	1	0	0	0	0	49
17	1984	0	8	0	2486	0	0	1	0	0	0	0	0	0	0	13	0	1.62	0	0	0	1	0	0	0	64
17	1985	0	9	0	2164	0	0	0	0	0	0	0	1	0	0	13	0	1.61	0	0	0	0	1	0	0	81
17	1986	0	10	0	2749	0	0	0	0	0	0	0	1	0	0	13	0	1.57	0	0	0	0	0	1	0	100
17	1987	0	11	0	2476	0	0	0	0	0	1	0	0	0	0	13	0	1.82	0	0	0	0	0	0	1	121
18	1980	0	4	0	2332	1	0	0	0	1	0	0	0	0	0	12	0	1.52	0	0	0	0	0	0	0	16
18	1981	0	5	0	2116	1	0	0	0	1	0	0	0	0	0	12	0	1.74	1	0	0	0	0	0	0	25
18	1982	0	6	0	2500	1	0	1	0	0	0	0	0	0	0	12	0	1.63	0	1	0	0	0	0	0	36
18	1983	0	7	0	2474	1	0	0	1	0	0	0	0	0	0	12	0	2.00	0	0	1	0	0	0	0	49
18	1984	0	8	0	2362	1	0	0	1	0	0	0	0	0	0	12	0	2.18	0	0	0	1	0	0	0	64
18	1985	0	9	0	2340	1	0	0	1	0	0	0	0	0	0	12	0	2.27	0	0	0	0	1	0	0	81
18	1986	0	10	0	2340	1	0	0	1	0	0	0	0	0	0	12	0	2.07	0	0	0	0	0	1	0	100
18	1987	0	11	0	2340	1	0	0	1	0	0	0	0	0	0	12	0	2.87	0	0	0	0	0	0	1	121
45	1980	0	2	0	1864	0	0	0	0	0	0	0	1	0	0	12	1	1.89	0	0	0	0	0	0	0	4
45	1981	0	3	0	2021	0	0	0	0	0	0	1	0	0	0	12	1	1.47	1	0	0	0	0	0	0	9
45	1982	0	4	0	2274	0	0	0	0	0	0	1	0	0	0	12	0	1.47	0	1	0	0	0	0	0	16
45	1983	0	5	0	2112	0	0	0	0	0	0	1	0	0	0	12	0	1.74	0	0	1	0	0	0	0	25
45	1984	0	6	0	1920	0	0	0	0	0	0	1	0	0	0	12	0	1.82	0	0	0	1	0	0	0	36

The first few lines of `wagepan2.tab`

30.3 Suggested exercise

1. To establish if the returns to education have changed over time we need to start by creating interaction effects for `educ` with the year dummy variables (`d81,d82,...,d87),` call these effects `edd81-edd97` respectively.
2. Estimate a homogeneous linear model for the response `lwage` with the covariates `espersq, union, married, d81-d87, edd81-edd97.` These results are given as the 1st part (homogeneous model) of the output that is obtained by estimating the fixed effects model. Estimate the model using the respondent fixed effects (`nr).` What is the main difference between the results from the alternative estimators?
3. Re-estimate the models of Task 2 without the time varying effects of education (`edd81-edd97).` Is the model a poorer fit to the data?
4. Re-estimate the fixed effects models of Task 2 using adaptive quadrature with `first.mass=12`. Compare the fixed and random effect model inferences. What do you find?

30.4 References

Vella, F., and Verbeek, M., (1998), Whose wages do unions raise? A dynamic model of unionism and wage rate determination for young men, Journal of Applied Econometrics, 13, 163-183.

Wooldridge, J. M. (2006), Introductory Econometrics: A Modern Approach. Third edition. Thompson, Australia.