Solutions Manual for sabreStata (Sabre in Stata) Exercises

Version 1 (Draft)

email: r.crouchley@lancaster.ac.uk

March 25, 2009

Abstract

Many users will have undertaken the exercises in interactive sessions. In this solutions manual we present the batch scripts that could be used to obtain the answers to the exercises. Sometimes the batch scripts are limited to the commands needed to obtain the last answer of the iterative model building and checking parts of the exercises, i.e. they do not include all the steps. Both the batch scripts, e.g. `grader.do` and their associated log files, e.g. `grader_s.log` are available from the Sabre site. Unless its otherwise made explicit in the text, when we use the term significant, we mean at the 95% level. It is also possible that we have failed to appreciate some of the complexities present in the data and covariates that are manifest in the many substantive fields from which these exercises are drawn, our apologies if this is the case

Contents

1 Exercise C1. Linear Model of Essay Grading

1.1 Relevant Results from grader_s.log and Discussion

Task 1. Estimate the linear model using Sabre on `grade`, with just a constant and no other effects.

Result/Discussion

```
Log likelihood =     -884.88956     on     394 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
cons                    5.2374          0.11374
sigma                   2.2635
```

Task 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.

Result/Discussion

```
Log likelihood =     -855.09330     on     393 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
cons                    5.2374          0.13958
sigma                   1.5827          0.79535E-01
scale                   1.6141          0.12628
```

The linear random effects model, only required 12 adaptive quadrature mass points. The `scale` parameter for this model suggests the presence of significant essay grade random effects.

Task 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.

Result/Discussion

```
Log likelihood =     -831.52131     on     392 residual degrees of freedom
```

```
Parameter              Estimate         Std. Err.
___________________________________________________
cons                    5.7525          0.15643
dg4                    -1.0303          0.14122
sigma                   1.4051          0.70609E-01
scale                   1.6943          0.11811
```

These results are for adaptive quadrature with 12 mass points.

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

Result/Discussion

The log likelihood of the homogeneous model of Task 1 is `-884.88956`, and log likelihood of the random effects model of Task 2 is `-855.09330`. The change in log likelihood over the homogeneous model is -2(-884.88956+855.09330)= 59. 593. The sampling distribution of this test statistic is not chi-square with 1 df. Under the null hypothesis `scale` has the value 0, and it can only take values >0 under the alternative. The correct p value for this test statistics is obtained by dividing the naive p value of 59. 593 for 1 degree of freedom by 1/2, and so its clearly significant, suggesting that the grades from the two graders are highly correlated. The log likelihood significantly reduces further when we add the grader indicator covariate `dg4.` This improvement in log likelihood has a chi-square of -2(-855.09330+ 831.52131)= 47. 144, for 1 more degree of freedom. The value of `scale` (sigma for the random effects) increases from `1.6141` in the model without covariates to `1.6943` for the model with the `dg4` indicator. The coefficient on `dg4` is negative `-1.0303 (S.E. 0.14122),` which is very significant, suggesting that grader 4 is a much lower marker than grader 1. All the estimated models assume a common sigma.

1.2 Batch Script: `grader.do`

```
log using grader_s.log, replace
set more off
use grader2
sabre, data ij r grade essay dg1 dg4
sabre ij r grade essay dg1 dg4, read
sabre, case essay
sabre, yvar grade
sabre, family g
sabre, constant cons
sabre, lfit cons
sabre, dis m
sabre, dis e
sabre, quad a
sabre, mass 12
sabre, fit cons
sabre, dis m
sabre, dis e
```

```
sabre, fit dg4 cons
sabre, dis m
sabre, dis e
log close
clear
exit
```

2 Exercise C2. Linear Model of Educational Attainment

2.1 Relevant Results from `neighborhood_s.log` and Discussion

Task 1. Estimate a linear model on attainment (`attain`) without covariates.

Result/Discussion

```
Log likelihood =     -3282.0735     on    2308 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
cons                   0.93396E-01      0.20850E-01
sigma                   1.0021
```

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

Result/Discussion

```
Number of observations            =    2310
Number of cases                   =      17

Log likelihood =     -3221.0818     on    2307 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
cons                   0.82269E-01      0.75715E-01
sigma                  0.96665          0.14279E-01
scale                  0.29790          0.58507E-01
```

The scale parameter estimate of 0.29790 (S.E. 0.58507E-01) has a z statistic of 0.29790/0.058507= 5.091 7, which is quite large, similarly with the associated change in log likelihood which has a chi-square of -2(-3282.0735+3221.0818)= 121.98.The sampling distribution of this test statistic is not chi-square with 1 df. Under the null hypothesis `scale` has the value 0, and it can only take values >0 under the alternative. The correct p value for this test statistics is obtained by dividing the naive p value of 121.98 for 1 degree of freedom by 1/2, and so its clearly significant.

Task 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?

Result/Discussion

```
Number of observations            =    2310
Number of cases                   =      17

Log likelihood =     -2403.9957     on    2300 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
cons                    0.80732E-01      0.26927E-01
p7vrq                   0.28319E-01      0.22811E-02
p7read                  0.27103E-01      0.17586E-02
dadocc                  0.94839E-02      0.13558E-02
dadunemp               -0.14941          0.46945E-01
daded                   0.15227          0.41103E-01
momed                   0.65025E-01      0.37709E-01
male                   -0.54138E-01      0.28642E-01
sigma                   0.68347          0.10094E-01
scale                   0.56053E-01      0.21390E-01
```

The scale parameter estimate shrinks from 0.29790 (S.E. 0.58507E-01) in the model without covariates to 0.56053E-01 (S.E. 0.21390E-01) for the model with the student specific effects.

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

Result/Discussion

```
Log likelihood =     -2384.8141     on    2299 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
cons                    0.85822E-01      0.27618E-01
p7vrq                   0.27557E-01      0.22644E-02
p7read                  0.26292E-01      0.17502E-02
dadocc                  0.81675E-02      0.13600E-02
dadunemp               -0.12076          0.46813E-01
daded                   0.14445          0.40787E-01
momed                   0.59444E-01      0.37394E-01
male                   -0.56061E-01      0.28403E-01
deprive                -0.15668          0.25269E-01
sigma                   0.67754          0.10004E-01
scale                   0.62311E-01      0.20628E-01
```

This model can be estimated with 12 adaptive quadrature mass points. The scale parameter estimate increases from 0.56053E-01 (S.E. 0.21390E-01) for the model with just the student specific effects to 0.62311E-01 (S.E. 0.20628E-01) for the model with the student specific effects and the neigbhourhood effect (`deprive`).

We now use a data set sorted by the neighbourhood identifier (`neighid`); called `neighbourhood2.dta`.

Task 5. Re-estimate the constant only model allowing for neighbourhood random effect (`neighid`), use adaptive quadrature with mass 12. Is there a significant `neighd` random effect?

Result/Discussion

The neighbourhood random effect (`neighid`) model with adaptive quadrature with mass 12 gives.

```
Log likelihood =     -3207.9848     on    2307 residual degrees of freedom

    Parameter               Estimate         Std. Err.
    ___________________________________________________
    cons                    0.82025E-01      0.28440E-01
    sigma                   0.89687          0.14815E-01
    scale                   0.44893          0.28651E-01
```

The associated change in log likelihood over the homogenous model of Task 1 has a chi-square of -2(-3282.0735+3207.9848)= 148. 18. The sampling distribution of this test statistic is not chi-square with 1 df. Under the null hypothesis `scale` has the value 0, and it can only take values >0 under the alternative. The correct p value for this test statistics is obtained by dividing the naive p value of 148. 18 for 1 degree of freedom by 1/2, and so its clearly significant.

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

Result/Discussion

```
    Log likelihood =     -2403.9492     on    2300 residual degrees of freedom

    Parameter               Estimate         Std. Err.
    ___________________________________________________
    cons                    0.77383E-01      0.23439E-01
    p7vrq                   0.28441E-01      0.22695E-02
    p7read                  0.26825E-01      0.17553E-02
    dadocc                  0.93107E-02      0.13681E-02
```

dadunemp	-0.14359	0.46900E-01
daded	0.14818	0.41109E-01
momed	0.67291E-01	0.37698E-01
male	-0.54457E-01	0.28608E-01
sigma	0.67583	0.11010E-01
scale	0.11593	0.31606E-01

The scale parameter estimate shrinks from 0.44893 (S.E. 0.28651E-01) in the model without covariates to 0.11593 (S.E. 0.31606E-01) for the model with the student specific effects.

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

Result/Discussion

```
Log likelihood =     -2387.4993     on    2299 residual degrees of freedom
```

Parameter	Estimate	Std. Err.
cons	0.80731E-01	0.22960E-01
p7vrq	0.27763E-01	0.22561E-02
p7read	0.26065E-01	0.17467E-02
dadocc	0.82389E-02	0.13668E-02
dadunemp	-0.11490	0.46832E-01
daded	0.14097	0.40829E-01
momed	0.62405E-01	0.37454E-01
male	-0.55381E-01	0.28434E-01
deprive	-0.14812	0.25331E-01
sigma	0.67574	0.11007E-01
scale	0.78917E-01	0.43246E-01

The scale parameter estimate increases from 0.11593 (S.E. 0.31606E-01) for the model with just the student specific effects to 0.78917E-01 (S.E. 0.43246E-01) for the model with the student specific effects and the neigbhourhood effect (`deprive`). The scale parameter in the model with student specific effects and the neigbhourhood effect is not significant, it has a z statistic 0.78917E-01/0.43246E-01$= -1.5232$.

Task 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?

Result/Discussion

Both the `schid` or the `neighid` random effects models are 2 level models, perhaps a 3 level model would be more appropriate on this data, i.e. pupils in schools, and schools in neigbhourhoods.

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

Result/Discussion

That both student specific and neigbhourhood effect (`deprive`) effects can be present in linear model of student attainment (`attain`). We can interpret the various covariate effects, e.g. the neigbhourhood effect (`deprive`) a measure of social deprivation has a very significant negative effect on student attainment.

2.2 Batch Script: `neighborhood.do`

```
log using neighborhood_s.log, replace
set more off
use neighborhood
#delimit ;
sabre, data neighid schid attain p7vrq p7read dadocc dadunemp daded momed
            male deprive dummy;
sabre neighid schid attain p7vrq p7read dadocc dadunemp daded momed male
      deprive dummy, read;
#delimit cr
sabre, case schid
sabre, yvar attain
sabre, family g
sabre, constant cons
sabre, lfit cons
sabre, dis m
sabre, dis e
sabre, quad a
sabre, mass 12
sabre, fit cons
sabre, dis m
sabre, dis e
sabre, fit p7vrq p7read dadocc dadunemp daded momed male cons
sabre, dis m
sabre, dis e
sabre, fit p7vrq p7read dadocc dadunemp daded momed male deprive cons
sabre, dis m
sabre, dis e
sort neighid
#delimit ;
sabre, data neighid schid attain p7vrq p7read dadocc dadunemp daded momed
            male deprive dummy;
sabre neighid schid attain p7vrq p7read dadocc dadunemp daded momed male
      deprive dummy, read;
#delimit cr
sabre, case neighid
sabre, yvar attain
sabre, family g
sabre, constant cons
sabre, quad a
sabre, mass 12
sabre, fit cons
sabre, dis m
sabre, dis e
sabre, fit p7vrq p7read dadocc dadunemp daded momed male cons
```

```
sabre, dis m
sabre, dis e
sabre, fit p7vrq p7read dadocc dadunemp daded momed male deprive cons
sabre, dis m
sabre, dis e
log close
clear
exit
```

3 Exercise C3. Binary Response Model of Essay Grades

3.1 Relevant Results from essays_s.log and Discussion

Task 1. Fit a binary probit model to the binary response `pass`, but without any random effects.

Result/Discussion

```
Log likelihood =     -686.20763     on     989 residual degrees of freedom

Parameter               Estimate         Std. Err.
___________________________________________________
cons                    0.50639E-02      0.39833E-01
```

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

Result/Discussion

```
Log likelihood =     -613.87204     on     988 residual degrees of freedom

Parameter               Estimate         Std. Err.
___________________________________________________
cons                    0.56694E-02      0.85207E-01
scale                   0.99151          0.95013E-01
```

The result above is for an 12 mass adaptive quadrature model, the essay random effect is significant, the change in log likelihood over the homogeneous model is -2(-686.20763+613.87204)= 144. 67. The sampling distribution of this test statistic is not chi-square with 1 df. Under the null hypothesis `scale` has the value 0, and it can only take values >0 under the alternative. The correct p value for this test statistics is obtained by dividing the naive p value of 144. 67 for 1 degree of freedom by 1/2, and so its clearly significant.

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

Result/Discussion

```
Log likelihood =     -562.68165     on     984 residual degrees of freedom

Parameter              Estimate        Std. Err.
---------------------------------------------------
cons                   0.86777         0.14749
grader2                -1.2153         0.16676
grader3               -0.72212         0.15941
grader4               -0.84969         0.16199
grader5                -1.5143         0.17153
scale                   1.1795         0.11237
```

All the grader indicator effects are negative, relative to grader1 (the reference category) and they all have significant t statistics. The estimated scale parameter and its standard error have increased slightly. Relative to grader1, the lowest marker is `grader5`, then we have `grader2, 4` and `3`.

Task 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?

Result/Discussion

```
Log likelihood =     -502.95053     on     978 residual degrees of freedom

Parameter              Estimate        Std. Err.
---------------------------------------------------
cons                   -6.8057          1.1242
grader2                -1.2084         0.16632
grader3               -0.71298         0.15895
grader4               -0.83704         0.16079
grader5                -1.5031         0.17052
wordlength              1.0244         0.23545
sqrtwords              0.29128         0.32422E-01
commas                 0.73205E-01     0.32721E-01
errors                -0.14654         0.39031E-01
prepos                 0.58790E-01     0.23941E-01
sentlength             0.35979E-03     0.12914E-01
scale                  0.71452         0.89305E-01
```

Only the `sentlength` essay characteristic is not significant in this extended model, `sqrtwords` is the most significant of the essay characteristics.

Task 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?

Result/Discussion

```
Log likelihood =     -496.55002     on     974 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
cons                    -6.8155          1.2189
grader2                 -2.1526         0.70353
grader3                 -1.9486         0.68342
grader4                -0.61700         0.63845
grader5                -0.73613         0.65089
wordlength               1.0592         0.24128
sqrtwords               0.27533         0.56617E-01
commas                  0.73714E-01     0.33381E-01
errors                 -0.14677         0.39805E-01
prepos                  0.59744E-01     0.24425E-01
sentlength              0.95757E-04     0.13170E-01
grader2sqrt             0.98148E-01     0.73308E-01
grader3sqrt             0.13425         0.73450E-01
grader4sqrt            -0.23209E-01     0.68602E-01
grader5sqrt            -0.77556E-01     0.68640E-01
scale                   0.73533         0.91437E-01
```

The model with interactions between the grader specific dummy variables and `sqrtwords` has a significant chi-square improvement of -2(-502.95053+496.55002)= -12.801 for 4 df. So there appears to be a different relationship between the length of the essay and essay grader for essay grade. However two of the grader indicators main effects i.e. `grader4`, `grader5`, have become non significant. The estimated scale parameter is still significant.

3.2 Batch Script: `essays.do`

```
log using essays_s.log, replace
set more off
use essays2
#delimit ;
sabre, data essay grader grade rating constant wordlength sqrtwords commas
            errors prepos sentlength pass grader2 grader3 grader4 grader5;
sabre essay grader grade rating constant wordlength sqrtwords commas errors
      prepos sentlength pass grader2 grader3 grader4 grader5, read;
#delimit cr
sabre, case essay
sabre, yvar pass
sabre, link p
```

```
sabre, constant cons
sabre, lfit cons
sabre, dis m
sabre, dis e
sabre, quad a
sabre, mass 12
sabre, fit cons
sabre, dis m
sabre, dis e
sabre, fit grader2 grader3 grader4 grader5 cons
sabre, dis m
sabre, dis e
#delimit ;
sabre, fit grader2 grader3 grader4 grader5 wordlength sqrtwords commas
           errors prepos sentlength cons;
#delimit cr
sabre, dis m
sabre, dis e
sabre, trans grader2sqrt grader2 * sqrtwords
sabre, trans grader3sqrt grader3 * sqrtwords
sabre, trans grader4sqrt grader4 * sqrtwords
sabre, trans grader5sqrt grader5 * sqrtwords
#delimit ;
sabre, fit grader2 grader3 grader4 grader5 wordlength sqrtwords commas
           errors prepos sentlength grader2sqrt grader3sqrt grader4sqrt
           grader5sqrt cons;
#delimit cr
sabre, dis m
sabre, dis e
log close
clear
exit
```

4 Exercise C4. Ordered Response Model of Essay Grades

4.1 Relevant Results from `essays_ordered_s.log` and Discussion

Task 1. Fit an ordered probit model to `ngrade` but without any random effects.

Result/Discussion

```
    Log likelihood =     -1371.6074     on     987 residual degrees of freedom

    Parameter              Estimate         Std. Err.
    ___________________________________________________
    cut1                  -0.66341         0.43188E-01
    cut2                  -0.50639E-02     0.39833E-01
    cut3                   0.62909         0.42834E-01
```

Task 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?

Result/Discussion

```
Log likelihood =     -1247.5966     on     986 residual degrees of freedom

    Parameter              Estimate         Std. Err.
    ___________________________________________________
    cut1                  -0.93258         0.89587E-01
    cut2                   0.24248E-02     0.85205E-01
    cut3                   0.88906         0.88940E-01
    scale                   1.0044         0.76825E-01
```

This model was estimated with 12 mass points. The change in log likelihood over the homogeneous model has a chi-square of -2(-1371.6074+1247.5966)= 248. 02. The sampling distribution of this test statistic is not chi-square with 1 df. Under the null hypothesis `scale` has the value 0, and it can only take values >0 under the alternative. The correct p value for this test statistics is obtained by dividing the naive p value of 248. 02 for 1 degree of freedom by 1/2, and so its clearly significant.

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

Result/Discussion

```
Log likelihood =     -1181.4489     on     982 residual degrees of freedom

Parameter              Estimate         Std. Err.
---------------------------------------------------
grader2                 -1.0885          0.12214
grader3                -0.63255          0.12004
grader4                -0.72804          0.11878
grader5                 -1.2842          0.12316
cut1                    -1.7957          0.13341
cut2                   -0.74225          0.12268
cut3                    0.25090          0.12080
scale                    1.1464          0.85246E-01
```

Relative to `grader1`, `grader5`, is the lowest marker followed by 2, 4 and 3.

Task 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?

Result/Discussion

```
Log likelihood =     -1116.1052     on     976 residual degrees of freedom

    Parameter              Estimate         Std. Err.
    ---------------------------------------------------
    grader2                 -1.0895          0.12193
    grader3                -0.62905          0.12001
    grader4                -0.72839          0.11846
    grader5                 -1.2849          0.12285
    wordlength              0.78477          0.20186
    sqrtwords               0.28050          0.26610E-01
    commas                  0.64009E-01      0.28346E-01
    errors                 -0.16114          0.33795E-01
    prepos                  0.50995E-01      0.20497E-01
    sentlength             -0.17035E-02      0.11399E-01
    cut1                    4.5615           0.93449
    cut2                    5.6071           0.93976
    cut3                    6.6058           0.94601
    scale                   0.71413          0.66264E-01
```

The covariate `sentlength` is not significant (z test). The change in log likelihood for adding the 6 essay characteristics is clearly significant, it has a chi-square of -2(-1181.4489+1116.1052)= 130.69.

Task 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?

Result/Discussion

```
Log likelihood =     -1094.4282     on     972 residual degrees of freedom

    Parameter              Estimate        Std. Err.
    -------------------------------------------------------
    grader2                 -1.3937         0.48887
    grader3                 -2.3223         0.51754
    grader4                 0.20938         0.46529
    grader5                -0.18398         0.47409
    wordlength              0.81793         0.20952
    sqrtwords               0.30176         0.43879E-01
    commas                  0.65559E-01     0.29393E-01
    errors                 -0.16543         0.35045E-01
    prepos                  0.52336E-01     0.21281E-01
    sentlength             -0.13918E-02     0.11819E-01
    grader2sqrt             0.28753E-01     0.51366E-01
    grader3sqrt             0.18273         0.55407E-01
    grader4sqrt            -0.10301         0.49563E-01
    grader5sqrt            -0.11935         0.50102E-01
    cut1                     4.8642          1.0185
    cut2                     5.9357          1.0234
    cut3                     6.9724          1.0301
    scale                   0.75099         0.68526E-01
```

The change in log likelihood has a chi-square of -2(-1116.1052+1094.4282)= 43.354 for 4 df, clearly significant overall. Various covariate effects are not significant in the model, these include `grader4, grader5, sentlength` and the interaction effect `grader2sqrt`.

Task 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.

Result/Discussion

```
Log likelihood =     -1707.3256     on     968 residual degrees of freedom

    Parameter              Estimate        Std. Err.
    -------------------------------------------------------
    grader2                 -1.3262         0.44038
    grader3                 -2.1009         0.45656
    grader4                 0.60237         0.42374
    grader5                 0.55202E-02     0.42951
    wordlength              0.90840         0.20927
    sqrtwords               0.33947         0.41241E-01
```

```
commas                  0.69427E-01      0.29487E-01
errors                 -0.15169          0.34760E-01
prepos                  0.54245E-01      0.20958E-01
sentlength              0.79695E-03      0.11814E-01
grader2sqrt             0.16434E-01      0.46235E-01
grader3sqrt             0.15843          0.48636E-01
grader4sqrt            -0.14085          0.44981E-01
grader5sqrt            -0.14256          0.45402E-01
cut1                    4.7135           1.0044
cut2                    5.6454           1.0062
cut3                    6.2119           1.0079
cut4                    6.7729           1.0104
cut5                    7.3627           1.0139
cut6                    7.8499           1.0170
cut7                    8.4523           1.0209
scale                   0.78548          0.62818E-01
```

We have only presented the result for full model with 7 cut points. Various covariate effects are not significant, these include `grader4, grader5, sentlength` and the interaction effect `grader2sqrt`.

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

Result/Discussion

If the model is correct the covariate parameter estimates should be similar from the model based on the 4 aggregate `ngrade` categories to those of the model based on the original 8 `grade` categories, as aggregation used in `ngrade` is of adjacent categories from `grade`. The ordered model using the 8 `grade` categories is to be preferred, as it contains more information about the ordered `grade`. This is generally true, so long as the response data are not too sparse across the categories. The cut points from the `grade` categories model suggest that the distance between `cut1` and `cut2`, (about 0.9) is greater than that between any other cut points (about 0.5). The `ngrade` and `grade` models agree about the covariates effects that are significant and non significant. There are small differences in the magnitude of the significant covariates, but they do not appear to be too large to suggest that there is a problem with the model.

4.2 Batch Script: `essays_ordered.do`

```
log using essays_ordered_s.log, replace
set more off
use essays_ordered
#delimit ;
sabre, data essay grader grade rating constant wordlength sqrtwords commas
            errors prepos sentlength pass grader2 grader3 grader4 grader5
            ngrade;
sabre essay grader grade rating constant wordlength sqrtwords commas errors
      prepos sentlength pass grader2 grader3 grader4 grader5 ngrade, read;
#delimit cr
sabre, case essay
```

```
sabre, yvar ngrade
sabre, ordered y
sabre, link p
sabre, lfit
sabre, dis m
sabre, dis e
sabre, quad a
sabre, mass 12
sabre, fit
sabre, dis m
sabre, dis e
sabre, fit grader2 grader3 grader4 grader5
sabre, dis m
sabre, dis e
#delimit ;
sabre, fit grader2 grader3 grader4 grader5 wordlength sqrtwords commas
           errors prepos sentlength;
#delimit cr
sabre, dis m
sabre, dis e
sabre, trans grader2sqrt grader2 * sqrtwords
sabre, trans grader3sqrt grader3 * sqrtwords
sabre, trans grader4sqrt grader4 * sqrtwords
sabre, trans grader5sqrt grader5 * sqrtwords
#delimit ;
sabre, fit grader2 grader3 grader4 grader5 wordlength sqrtwords commas
           errors prepos sentlength grader2sqrt grader3sqrt grader4sqrt
           grader5sqrt;
#delimit cr
sabre, dis m
sabre, dis e
sabre, yvar grade
sabre, lfit
sabre, dis m
sabre, dis e
sabre, quad a
sabre, mass 12
sabre, fit
sabre, dis m
sabre, dis e
sabre, fit grader2 grader3 grader4 grader5
sabre, dis m
sabre, dis e
#delimit ;
sabre, fit grader2 grader3 grader4 grader5 wordlength sqrtwords commas
           errors prepos sentlength;
#delimit cr
sabre, dis m
sabre, dis e
#delimit ;
sabre, fit grader2 grader3 grader4 grader5 wordlength sqrtwords commas
           errors prepos sentlength grader2sqrt grader3sqrt grader4sqrt
           grader5sqrt;
#delimit cr
sabre, dis m
sabre, dis e
log close
clear
exit
```

5 Exercise C5. Poison Model of Headaches

5.1 Relevant Results from headache2_s.log and Discussion

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

Result/Discussion

```
trans lt log days
```

Task 2. Fit a Poisson model to y (number of headaches) with a log link without any id random effects.

Result/Discussion

```
Log likelihood =     -234.50796     on     120 residual degrees of freedom

    Parameter              Estimate         Std. Err.
    ---------------------------------------------------
    cons                   -1.3972          0.69843E-01
```

Task 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?

Result/Discussion

```
    Log likelihood =     -205.61598     on     120 residual degrees of freedom

    Parameter              Estimate         Std. Err.
    ---------------------------------------------------
    cons                   -1.6035          0.15971
    scale                  0.68943          0.13888
```

We used 12 adaptive quadrature points. This gave a chi-square improvement of -2(-234.50796+205.61598)= 57.784.over the homogeneous model. The sampling distribution of this test statistic is not chi-square with 1 df. Under the null hypothesis scale has the value 0, and it can only take values >0 under the alternative. The correct p value for this test statistics is obtained by dividing the naive p value of 57.784 for 1 degree of freedom by 1/2, and so its clearly significant.

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

Result/Discussion

```
Log likelihood =     -203.66800     on     119 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
cons                   -1.7154          0.17187
aspartame              0.28246          0.14216
scale                  0.69543          0.14002
```

The treatment indicator `aspartame` has a significant z statistic, its 0.28246/0.14002= 2.0173.

5.2 Batch Script: `headache2.do`

```
log using headache2_s.log, replace
set more off
use headache2
sabre, data id y constant aspartame days
sabre id y constant aspartame days, read
sabre, case id
sabre, yvar y
sabre, family p
sabre, constant cons
sabre, trans lt log days
sabre, offset lt
sabre, lfit cons
sabre, dis m
sabre, dis e
sabre, quad a
sabre, mass 12
sabre, fit cons
sabre, dis m
sabre, dis e
sabre, fit aspartame cons
sabre, dis m
sabre, dis e
log close
clear
exit
```

6 Exercise L1. Linear Model of Psychological Distress

6.1 Relevant Results from ghq_s.log and Discussion

Task 1. Estimate the linear model in sabre on ghq, with just a constant, and no random effects.

Result/Discussion

```
Log likelihood =     -76.935774     on      22 residual degrees of freedom

Parameter              Estimate        Std. Err.
___________________________________________________
cons                    10.167          1.2448
sigma                   6.0982
```

Task 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?

Result/Discussion

```
Log likelihood =     -67.132857     on      21 residual degrees of freedom

Parameter              Estimate        Std. Err.
___________________________________________________
cons                    10.167          1.6784
sigma                   1.9149         0.39087
scale                   5.6544          1.2222
```

The change in log likelihood over the homogeneous model has a chi-square of -2(-76.935774+67.132857). = 19.606 The sampling distribution of this test statistic is not chi-square with 1 df. Under the null hypothesis scale has the value 0, and it can only take values >0 under the alternative. The correct p value for this test statistics is obtained by dividing the naive p value of 19.606 for 1 degree of freedom by 1/2, and so its clearly significant.

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

Result/Discussion

```
Log likelihood =     -67.041252     on      20 residual degrees of freedom

    Parameter              Estimate         Std. Err.
    ___________________________________________________
    cons                     10.333          1.7227
    dg2                    -0.33333         0.77579
    sigma                    1.9003         0.38789
    scale                    5.6568          1.2216
```

The change in log likelihood has a chi-square of -2(-67.132857+67.041252)= 0.183 21 for 1 df, which is not significant. The z statistic for the dg2 estimate is -0.33333/0.77579= $-0.429\,67$, which is also non significant. These results imply that there is no occasion effect on psychological distress in the data.

6.2 Batch Script: ghq.do

```
log using ghq_s.log, replace
set more off
use ghq2
sabre, data ij r student ghq dg1 dg2
sabre ij r student ghq dg1 dg2, read
sabre, case student
sabre, yvar ghq
sabre, family g
sabre, constant cons
sabre, lfit cons
sabre, dis m
sabre, dis e
sabre, quad a
sabre, mass 12
sabre, fit cons
sabre, dis m
sabre, dis e
sabre, fit dg2 cons
sabre, dis m
sabre, dis e
log close
clear
exit
```

7 Exercise L2. Linear Model of log Wages

7.1 Relevant Results from wagepan_s.log and Discussion

Task 1. Estimate a linear model on lwage (log of hourly wage) without covariates.

Result/Discussion

```
Log likelihood =     -3439.4161     on    4358 residual degrees of freedom

Parameter              Estimate         Std. Err.
---------------------------------------------------
cons                     1.6491         0.80661E-02
sigma                   0.53261
```

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

Result/Discussion

```
Log likelihood =     -2621.1724     on    4357 residual degrees of freedom

    Parameter              Estimate         Std. Err.
    ---------------------------------------------------
    cons                     1.6491         0.16722E-01
    sigma                   0.38723         0.44331E-02
    scale                   0.36559         0.12640E-01
```

This model has a chi-square improvement of -2(-3439.4161+2621.1724)= 1636.5.over the homogeneous model. The sampling distribution of this test statistic is not chi-square with 1 df. Under the null hypothesis scale has the value 0, and it can only take values >0 under the alternative. The correct p value for this test statistics is obtained by dividing the naive p value of 1636.5 for 1 degree of freedom by 1/2, and so its clearly significant.

Task 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?

Result/Discussion

```
Log likelihood =     -2186.9588     on    4343 residual degrees of freedom

Parameter              Estimate         Std. Err.
---------------------------------------------------
```

```
cons                     0.23164E-01      0.15233
educ                     0.91887E-01      0.10780E-01
black                   -0.13938          0.48258E-01
hisp                     0.21774E-01      0.43089E-01
exper                    0.10598          0.15445E-01
expersq                 -0.47369E-02      0.68805E-03
married                  0.63565E-01      0.16779E-01
union                    0.10548          0.17885E-01
fyear       (    1)       0.0000          ALIASED [I]
fyear       (    2)      0.40367E-01      0.24682E-01
fyear       (    3)      0.30749E-01      0.32458E-01
fyear       (    4)      0.20054E-01      0.41838E-01
fyear       (    5)      0.42859E-01      0.51713E-01
fyear       (    6)      0.57522E-01      0.61771E-01
fyear       (    7)      0.91653E-01      0.71910E-01
fyear       (    8)      0.13470          0.82135E-01
sigma                    0.35066          0.40172E-02
scale                    0.32987          0.11470E-01
```

This gave a chi-square improvement of -2(-2621.1724+2186.9588)= 868.43 for 4357-4343= 14 df, which is very significant overall. But judged by the various covariate parameter estimates the following main effects are not significant: `hisp`, `fyear(2-7)`. The `scale` parameter in the model with covariates is slightly smaller.

Task 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?

Result/Discussion

```
Log likelihood =     -2185.7569     on    4336 residual degrees of freedom

    Parameter                Estimate         Std. Err.

    ---------------------------------------------------
    cons                    -0.30601E-01      0.18810
    educ                     0.94647E-01      0.13702E-01
    black                   -0.13961          0.48306E-01
    hisp                     0.22405E-01      0.43134E-01
    exper                    0.11554          0.17029E-01
    expersq                 -0.53658E-02      0.83374E-03
    married                  0.64033E-01      0.16782E-01
    union                    0.10448          0.17895E-01
    fyear       (    1)       0.0000          ALIASED [I]
    fyear       (    2)     -0.28781E-01      0.14519
    fyear       (    3)     -0.10056E-01      0.14673
    fyear       (    4)      0.17697E-01      0.14949
```

```
fyear       (    5)      0.11328            0.15367
fyear       (    6)      0.11713            0.15942
fyear       (    7)      0.17924            0.16686
fyear       (    8)      0.25606            0.17614
educ81                   0.54357E-02        0.12197E-01
educ82                   0.26951E-02        0.12298E-01
educ83                  -0.79957E-03        0.12466E-01
educ84                  -0.71021E-02        0.12700E-01
educ85                  -0.61964E-02        0.12992E-01
educ86                  -0.84785E-02        0.13339E-01
educ87                  -0.11141E-01        0.13741E-01
sigma                    0.35051            0.40155E-02
scale                    0.33026            0.11483E-01
```

The addition of the interaction effects gave a chi-square improvement of -2(-2186.9588+2185.7569)= 2.4038 for 4343-4336= 7 df, which is not significant. None of the individual interaction effects have significant z statistics, i.e. returns to education do not appear to change with year. Both the interaction effects and the main effects of `year` could be removed from this model. The `scale` parameter is still significant, suggesting a correlation between log wages for an individual over successive years.

7.2 Batch Script: `wagepan.do`

```
log using wagepan_s.log, replace
set more off
use wagepan
#delimit ;
sabre, data nr year agric black bus construc ent exper fin hisp poorhlth
            hours manuf married min nrthcen nrtheast occ1 occ2 occ3 occ4
            occ5 occ6 occ7 occ8 occ9 per pro pub rur south educ tra trad
            union lwage d81 d82 d83 d84 d85 d86 d87 expersq;
sabre nr year agric black bus construc ent exper fin hisp poorhlth hours
      manuf married min nrthcen nrtheast occ1 occ2 occ3 occ4 occ5 occ6 occ7
      occ8 occ9 per pro pub rur south educ tra trad union lwage d81 d82 d83
      d84 d85 d86 d87 expersq, read;
#delimit cr
sabre, case nr
sabre, yvar lwage
sabre, family g
sabre, constant cons
sabre, fac year fyear
sabre, lfit cons
sabre, dis m
sabre, dis e
sabre, quad a
sabre, mass 12
sabre, fit cons
sabre, dis m
sabre, dis e
sabre, fit educ black hisp exper expersq married union fyear cons
sabre, dis m
sabre, dis e
sabre, trans educ81 educ * d81
sabre, trans educ82 educ * d82
sabre, trans educ83 educ * d83
```

```
sabre, trans educ84 educ * d84
sabre, trans educ85 educ * d85
sabre, trans educ86 educ * d86
sabre, trans educ87 educ * d87
#delimit ;
sabre, fit educ black hisp exper expersq married union fyear educ81 educ82
           educ83 educ84 educ85 educ86 educ87 cons;
#delimit cr
sabre, dis m
sabre, dis e
log close
clear
exit
```

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

8.1 Relevant Results from `ezunem_s.log` and Discussion

Task 1. Estimate a linear model on the log of number of unemployment claims (`luclms`) without covariates.

Result/Discussion

```
Log likelihood =     -213.81328     on     196 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
cons                    11.191          0.50759E-01
sigma                  0.71424
```

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

Result/Discussion

```
Log likelihood =     -166.35513     on     195 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
cons                    11.191          0.11550
sigma                  0.49075          0.26157E-01
scale                  0.51645          0.85713E-01
```

This model has a chi-square improvement of -2(-213.81328+166.35513)= 94. 916 over the homogeneous model. The sampling distribution of this test statistic is not chi-square with 1 df. Under the null hypothesis `scale` has the value 0, and it can only take values >0 under the alternative. The correct p value for this test statistics is obtained by dividing the naive p value of 94. 916 for 1 degree of freedom by 1/2, and so its clearly significant.

Task 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?

Result/Discussion

```
Log likelihood =     -135.33303     on     194 residual degrees of freedom

Parameter               Estimate         Std. Err.
---------------------------------------------------
cons                      11.363         0.12453
ez                      -0.74164         0.85576E-01
sigma                    0.40825         0.21770E-01
scale                    0.56033         0.89814E-01
```

The `scale` parameter estimate is slightly larger in the model with the `ez` covariate. The `ez` parameter estimate has a z statistics of -0.74164/0.085576= -8.6664 which is clearly significant. The negative coefficient on `ez` suggests that the log of the number of unemployment claims is smaller in cites which are in the enterprise zone.

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

Result/Discussion

```
Log likelihood =     -59.438419     on     193 residual degrees of freedom

Parameter               Estimate         Std. Err.
---------------------------------------------------
cons                      11.918         0.12196
ez                      -0.13846         0.69012E-01
t                       -0.13906         0.90240E-02
sigma                    0.26722         0.14243E-01
scale                    0.53601         0.83053E-01
```

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

Result/Discussion

The Task 4 model has a chi-square improvement of -2(-135.33303+ 59.438419)= 151.79 over the Task 3 model. The `scale` parameter estimate is slightly smaller in the Task 4 model. Both the `ez` and `t` parameter estimates have significant z statistics. The magnitude of the negative `ez` parameter estimate in the Task 4 model is smaller than that of the Task 3 model. The coefficient on time `t` is negative, suggesting that both the enterprise and non enterprize zone unemployment claims are declining with year (1980-1988). The negative coefficient on `ez` suggests that the log of the number of unemployment claims is smaller in cites which are in the enterprise zone.

8.2 Batch Script: `ezunem.do`

```
log using ezunem_s.log, replace
set more off
use ezunem2
#delimit ;
sabre, data year uclms ez d81 d82 d83 d84 d85 d86 d87 d88 c1 c2 c3 c4 c5 c6
            c7 c8 c9 c10 c11 c12 c13 c14 c15 c16 c17 c18 c19 c20 c21 c22
            luclms t ezt city;
sabre year uclms ez d81 d82 d83 d84 d85 d86 d87 d88 c1 c2 c3 c4 c5 c6 c7 c8
      c9 c10 c11 c12 c13 c14 c15 c16 c17 c18 c19 c20 c21 c22 luclms t ezt
      city, read;
#delimit cr
sabre, case city
sabre, yvar luclms
sabre, family g
sabre, constant cons
sabre, lfit cons
sabre, dis m
sabre, dis e
sabre, quad a
sabre, mass 12
sabre, fit cons
sabre, dis m
sabre, dis e
sabre, fit ez cons
sabre, dis m
sabre, dis e
sabre, fit ez t cons
sabre, dis m
sabre, dis e
log close
clear
exit
```

9 Exercise L4. Binary Model of Trade Union Membership

9.1 Relevant Results from unionpan_s.log and Discussion

Task 1. Estimate a logit model for trade union membership (`union`), without covariates.

Result/Discussion

```
Log likelihood =     -2422.8016     on    4359 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
cons                    -1.1307         0.35260E-01
```

Task 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?

Result/Discussion

```
Log likelihood =     -1671.6755     on    4358 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
cons                    -2.4630         0.17429
scale                    3.0758         0.18129
```

This is the result with 72 adaptive quadrature mass points. This model has a a chi-square improvement of -2(-2422.8016+1671.6755)= 1502. 3 over the homogeneous model. The sampling distribution of this test statistic is not chi-square with 1 df. Under the null hypothesis `scale` has the value 0, and it can only take values >0 under the alternative. The correct p value for this test statistics is obtained by dividing the niave p value of 1502. 3 for 1 degree of freedom by 1/2, and so its clearly significant.

Task 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?

Result/Discussion

```
Log likelihood =     -1659.5364     on    4352 residual degrees of freedom

Parameter              Estimate        Std. Err.
___________________________________________________
cons                    -1.9169          1.1417
black                    1.7662         0.46632
hisp                    0.82086         0.42208
exper                  -0.45506E-01     0.24070E-01
educ                   -0.62424E-01     0.92438E-01
poorhlth               -0.75160         0.50254
married                 0.34208         0.15907
scale                    3.0203         0.17834
```

This gave a chi-square improvement of -2(-1671.6755+1659.5364)= 24.278 for 4358-4352= 6 df, which is very significant overall. But judged by the various covariate parameter estimates, the following main effects are not significant: `educ, poorhlth`, while `exper` has borderline significance. The `scale` parameter in the model with covariates is still very significant and only slightly smaller. This model suggests that respondents who are `black` or `hisp` are more likely to be trade union members than whites. It also suggests that workers with longer labour market experience (`exper`) are less likely to be trade union members. While those who are `married` are more likely to be trade union members.

Task 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?

Result/Discussion

```
Log likelihood =     -1654.9281     on    4348 residual degrees of freedom

Parameter              Estimate        Std. Err.
___________________________________________________
cons                    -2.4347          1.2006
black                    1.8870         0.47315
hisp                     1.1052         0.44739
exper                  -0.40595E-01     0.24199E-01
educ                   -0.60500E-01     0.93061E-01
poorhlth               -0.75608         0.50335
married                 0.34500         0.15984
rur                     0.20794         0.24023
nrthcen                 0.69825         0.38780
nrtheast                0.87514         0.42444
south                   0.31154E-01     0.38514
scale                    3.0130         0.17885
```

This gave a chi-square improvement over the model of Task 3 of -2(-1659.5364 +1654.9281)= 9.2166 for 4352-4348= 4 df, which is of marginal significant. But judging by the various covariate parameter estimates, the following contextual effects are not significant: `rur, south`, while `nrthcen` is of marginal significance. The `scale` parameter in the model with covariates is slightly smaller.

Task 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?

Result/Discussion

```
Log likelihood =     -1648.5200     on    4341 residual degrees of freedom

    Parameter              Estimate         Std. Err.
    ___________________________________________________
    cons                    -3.5267           1.5875
    black                    1.8547          0.47558
    hisp                     1.0994          0.44857
    exper                   0.78144E-01      0.11319
    educ                    0.29340E-02      0.11073
    poorhlth               -0.75088          0.50414
    married                 0.35840          0.16124
    rur                     0.16395          0.24218
    nrthcen                 0.69374          0.38903
    nrtheast                0.89547          0.42624
    south                   0.49953E-01      0.38611
    d81                    -0.13844          0.23405
    d82                    -0.14765          0.30445
    d83                    -0.37875          0.39646
    d84                    -0.40806          0.49582
    d85                    -0.81673          0.60154
    d86                     -1.0608          0.70928
    d87                    -0.55944          0.81502
    scale                    3.0219          0.17944
```

This gave a chi-square improvement of -2(-1654.9281+1648.5200)= 12.816 for 4348-4341= 7 df, which is not significant at the 0.05 level. This is backed up by the year dummy variable parameter estimates, as none of them are significant.

Task 6. Include interaction effects between `rur` and `nrthcen, nrtheast`, `south` and add them to the model. Are any of these new effects significant?

Result/Discussion

```
    Log likelihood =     -1646.0610     on    4338 residual degrees of freedom

    Parameter              Estimate         Std. Err.
    ___________________________________________________
    cons                    -3.5764           1.5937
```

black	1.8663	0.47779
hisp	1.1461	0.45152
exper	0.73943E-01	0.11369
educ	0.13740E-01	0.11129
poorhlth	-0.77703	0.50835
married	0.35646	0.16168
rur	-0.83058	0.73415
nrthcen	0.60996	0.40177
nrtheast	0.91324	0.43866
south	-0.18017	0.40372
d81	-0.13588	0.23467
d82	-0.14725	0.30549
d83	-0.36793	0.39818
d84	-0.39353	0.49811
d85	-0.79952	0.60429
d86	-1.0401	0.71255
d87	-0.53695	0.81878
rur_nrthcen	1.0602	0.85693
rur_nrtheast	0.32601	0.94706
rur_south	1.4901	0.82363
scale	3.0350	0.18135

This gave a chi-square improvement of -2(-1648.5200+1646.0610)= 4. 918 for 4341-4338=3 df, which is not significant at the 0.05 level.

Task 7. How can the final model be simplified?

Result/Discussion

We could drop some of the contextual covariates from the model, namely: the interaction effects between `rur` and `nrthcen, nrtheast`, `south and` the main effects of : `d81-d87, rur`, and `south`. We could also drop the individual specific covariates `exper, educ` and `poorhlth.`

Task 8. Interpret your preferred model.

Result/Discussion

The preferred model is that of Task 4. This model suggests that respondents who are `black` or `hisp` are more likely to be trade union members than whites. It also suggests that workers with longer labour market experience (`exper`) are less likely to be trade union members. While those who are `married` are more likely to be trade union members. Furthermore the respondents from `nrthcen` and the `nrtheast` US are more likely to be trade union members than the rest.

9.2 Batch Script: `unionpan.do`

```
log using unionpan_s.log, replace
set more off
use wagepan
```

```
#delimit ;
sabre, data nr year agric black bus construc ent exper fin hisp poorhlth
            hours manuf married min nrthcen nrtheast occ1 occ2 occ3 occ4
            occ5 occ6 occ7 occ8 occ9 per pro pub rur south educ tra trad
            union lwage d81 d82 d83 d84 d85 d86 d87 expersq;
sabre nr year agric black bus construc ent exper fin hisp poorhlth hours
      manuf married min nrthcen nrtheast occ1 occ2 occ3 occ4 occ5 occ6 occ7
      occ8 occ9 per pro pub rur south educ tra trad union lwage d81 d82 d83
      d84 d85 d86 d87 expersq, read;
#delimit cr
sabre, case nr
sabre, yvar union
sabre, constant cons
sabre, lfit cons
sabre, dis m
sabre, dis e
sabre, quad a
sabre, mass 72
sabre, fit cons
sabre, dis m
sabre, dis e
sabre, fit black hisp exper educ poorhlth married cons
sabre, dis m
sabre, dis e
#delimit ;
sabre, fit black hisp exper educ poorhlth married rur nrthcen nrtheast
           south cons;
#delimit cr
sabre, dis m
sabre, dis e
#delimit ;
sabre, fit black hisp exper educ poorhlth married rur nrthcen nrtheast south
           d81 d82 d83 d84 d85 d86 d87 cons;
#delimit cr
sabre, dis m
sabre, dis e
sabre, trans rur_nrthcen rur * nrthcen
sabre, trans rur_nrtheast rur * nrtheast
sabre, trans rur_south rur * south
#delimit ;
sabre, fit black hisp exper educ poorhlth married rur nrthcen nrtheast south
       d81 d82 d83 d84 d85 d86 d87 rur_nrthcen rur_nrtheast rur_south cons;
#delimit cr
sabre, dis m
sabre, dis e
log close
clear
exit
```

10 Exercise L5. Ordered Response Model of Attitudes to Abortion

10.1 Relevant Results from `abortion_s.log` and Discussion

Task 1. Estimate an ordered logit model to `nscore`, without covariates.

Result/Discussion

```
Log likelihood =     -1766.6663     on    1051 residual degrees of freedom

    Parameter               Estimate         Std. Err.
    ___________________________________________________
    cut1                     -2.5150         0.11697
    cut2                    -0.80171         0.66557E-01
    cut3                    -0.28216         0.62159E-01
    cut4                     0.18996         0.61824E-01
    cut5                     0.75342         0.65965E-01
```

Task 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?

Result/Discussion

```
Log likelihood =     -1556.6472     on    1050 residual degrees of freedom

    Parameter               Estimate         Std. Err.
    ___________________________________________________
    cut1                     -4.2791         0.24225
    cut2                     -1.4925         0.17958
    cut3                    -0.55745         0.17319
    cut4                     0.33330         0.17198
    cut5                      1.3759         0.17696
    scale                     2.4006         0.16334
```

This is the result with 24 adaptive quadrature mass points. This person level model has a chi-square improvement of -2(-1766.6663+1556.6472)= 420.04 over the homogeneous model. The sampling distribution of this test statistic is not chi-square with 1 df. Under the null hypothesis `scale` has the value 0, and it can only take values >0 under the alternative. The correct p value for this test statistics is obtained by dividing the naive p value of 420.04 for 1 degree of freedom by 1/2, and so its clearly significant.

Task 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?

Result/Discussion

```
Log likelihood =     -1540.5327     on    1039 residual degrees of freedom

    Parameter              Estimate        Std. Err.
    ___________________________________________________
    male                    0.16982         0.31372
    age                     0.95699E-03     0.10287E-01
    dr2                      1.8853         0.64382
    dr3                     0.55578         0.69683
    dr4                      2.6697         0.65074
    dp2                     0.12500         0.29870
    dp3                     0.64082E-01     0.30195
    dp4                    -0.10560E-01     0.51927
    dp5                    -0.20071E-01     0.56075
    dc2                    -0.27901         0.26781
    dc3                    -0.16280         0.27664
    cut1                    -2.4638         0.80401
    cut2                    0.33189         0.79843
    cut3                     1.2665         0.79976
    cut4                     2.1551         0.80193
    cut5                     3.1958         0.80568
    scale                    2.2332         0.15515
```

This gave a chi-square improvement of -2(-1556.6472+1540.5327)= 32.229 for 1050-1039= 11 df, which is significant at the 0.05 level. But judged by the various covariate parameter estimates, the following main effects are not significant: `male, age, dr3` (other religion)`,` the way the respondent votes (`dp2-5`)`,` and the respondent's self asses social class (`dc2-3`)`.` The `scale` parameter in the model with covariates is still very significant and only slightly smaller. This model, which is clustered by person over time, suggests that respondent's who are protestant (`dr2`) or agnostic (`dr4`) are more likely to support legalising abortion, and that other effects: e.g. gender, age, the way the respondent votes and their self assessed social class have no effect.

Task 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.dta`.

Result/Discussion

For the model without covariates we have

```
Log likelihood =     -1741.0190     on    1050 residual degrees of freedom
```

```
Parameter              Estimate        Std. Err.
__________________________________________________
cut1                    -2.6736          0.15065
cut2                   -0.89016          0.11418
cut3                   -0.33529          0.11119
cut4                    0.17788          0.11086
cut5                    0.79479          0.11360
scale                   0.64059          0.98315E-01
```

For the model with covariates we have

```
Log likelihood =     -1685.2618     on    1039 residual degrees of freedom

    Parameter              Estimate        Std. Err.
    __________________________________________________
    male                    0.21814          0.12569
    age                    -0.20262E-02      0.42416E-02
    dr2                     0.83663          0.26861
    dr3                    -0.70121E-01      0.29630
    dr4                      1.6493          0.26835
    dp2                     0.38519E-01      0.15562
    dp3                     0.33915E-01      0.16789
    dp4                    -0.18177          0.34109
    dp5                     0.19365          0.41731
    dc2                    -0.28431          0.17290
    dc3                    -0.31155          0.16514
    cut1                    -2.1957          0.37683
    cut2                   -0.29027          0.36705
    cut3                    0.31419          0.36646
    cut4                    0.87675          0.36628
    cut5                     1.5488          0.36737
    scale                   0.81142          0.11553
```

The results for the respondents clustered by district and over time are with 12 adaptive quadrature mass points. This gave a chi-square improvement of -2(-1741.0190+1685.2618)= 111.51 for 1050-1039= 11 df, which is significant at the 0.05 level. But judged by the various covariate parameter estimates, the following main effects are not significant: `male, age, dr3` (other religion), the way the respondent votes (`dp2-5),` and the respondent's self asses social class (`dc2-3).` The `scale` parameter in the district model with covariates is still very significant and larger than the value obtained from the district model without covariates. This model is clustered by district and thus includes persons over time suggests that respondent's who are protestant (`dr2`) or agnostic (`dr4`) are more likely to support legalising abortion, but that gender, age and the way the respondent votes and their self asses social class have no effect.

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

Result/Discussion

The covariate inferences for the person and district level models are very similar. The main difference is in the magnitude of the significant covariate effects, this occurs because of differences in the magnitude of the scale parameter. The magnitude of the scale parameter has an effect on the magnitude of the covariate effects in this class of ordered response models. The person level model has a `scale` of 2.4006 (S.E. 0.16334), while that of the district level model has a `scale` of 0.81142 (S.E. 0.11553).

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

Result/Discussion

While the district level effect includes the highly correlated responses of an individual over time, it also includes the low correlated responses of different individuals in the same district. Perhaps a 3 level model of time, respondents and districts with just the respondents religion as a covariate would be more appropriate.

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

Result/Discussion

It may be worth trying the 3 way interaction of religion with age and gender and including the associated two way interaction effects. It could be that respondent's become more conservative as they grow older, and the magnitude of this change could be different for men and women.

10.2 Batch Script: `abortion.do`

```
log using abortion_s.log, replace
set more off
use abortion2
#delimit ;
sabre, data district person year score age male nscore dr2 dr3 dr4 dp2 dp3
           dp4 dp5 dc2 dc3;
sabre district person year score age male nscore dr2 dr3 dr4 dp2 dp3 dp4 dp5
      dc2 dc3, read;
#delimit cr
sabre, case person
sabre, yvar nscore
sabre, ordered y
sabre, lfit
sabre, dis m
sabre, dis e
sabre, quad a
sabre, mass 24
sabre, fit
sabre, dis m
sabre, dis e
sabre, fit male age dr2 dr3 dr4 dp2 dp3 dp4 dp5 dc2 dc3
sabre, dis m
```

```
sabre, dis e
sort district
#delimit ;
sabre, data district person year score age male nscore dr2 dr3 dr4 dp2 dp3
            dp4 dp5 dc2 dc3;
sabre district person year score age male nscore dr2 dr3 dr4 dp2 dp3 dp4 dp5
      dc2 dc3, read;
#delimit cr
sabre, case district
sabre, yvar nscore
sabre, ordered y
sabre, quad g
sabre, quad a
sabre, mass 12
sabre, fit
sabre, dis m
sabre, dis e
sabre, fit male age dr2 dr3 dr4 dp2 dp3 dp4 dp5 dc2 dc3
sabre, dis m
sabre, dis e
log close
clear
exit
```

11 Exercise L6. Ordered Response Model of Respiratory Status

11.1 Relevant Results from respiratory_s.log and Discussion

Task 1. Estimate an ordered logit model for status without any covariates.

Result/Discussion

```
Log likelihood =     -829.79872     on     551 residual degrees of freedom

    Parameter              Estimate         Std. Err.
    ___________________________________________________
    cut1                    -2.4771         0.15877
    cut2                    -1.4790         0.10918
    cut3                   -0.14802         0.85128E-01
    cut4                    0.81744         0.92086E-01
```

Task 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?

Result/Discussion

```
Log likelihood =     -714.06206     on     550 residual degrees of freedom

    Parameter              Estimate         Std. Err.
    ___________________________________________________
    cut1                    -4.1063         0.32842
    cut2                    -2.5515         0.27634
    cut3                   -0.25255         0.24631
    cut4                     1.4646         0.25333
    scale                    2.2652         0.21966
```

This is the result with 20 adaptive quadrature mass points. This model has a chi-square improvement of -2(-829.79872+714.06206)= 231.47 over the homogeneous model. The sampling distribution of this test statistic is not chi-square with 1 df. Under the null hypothesis `scale` has the value 0, and it can only take values >0 under the alternative. The correct p value for this test statistics is obtained by dividing the naive p value of 231.47 for 1 degree of freedom by 1/2, and so its clearly significant.

Task 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?

Result/Discussion

```
Log likelihood =     -703.29855     on     546 residual degrees of freedom

Parameter              Estimate        Std. Err.
---------------------------------------------------
drug                    -1.4348        0.43353
male                   -0.30416        0.55166
age                    -0.16700E-01    0.16112E-01
base                    0.27552        0.81994E-01
cut1                    -6.5127         1.1787
cut2                    -4.9909         1.1554
cut3                    -2.7151         1.1349
cut4                   -0.98493         1.1264
scale                    1.9823        0.20691
```

This gave a chi-square improvement of -2(-714.06206+703.29855)= 21.527 for 550-546= 4 df, which is significant at the 0.05 level. But judged by the various covariate parameter estimates, the following main effects are not significant: `male, age.` The `scale` parameter in the model with covariates is still very significant and a little smaller. This model for respiratory status, which is clustered by respondent over visit, suggests that respondent's who are in the treatment group (`drug`) have a poorer response than those who were given the placebo, while those who had a high baseline response (`base`) are more likely to have a high respiratory response.

Task 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?

Result/Discussion

```
Log likelihood =     -703.02730     on     545 residual degrees of freedom

Parameter              Estimate        Std. Err.
---------------------------------------------------
drug                    -1.4317        0.42726
male                   -0.30837        0.54353
age                    -0.16850E-01    0.15877E-01
base                    0.32628        0.10685
trend                  -0.57596E-01    0.78104E-01
cut1                    -6.5221         1.1631
cut2                    -5.0032         1.1396
cut3                    -2.7355         1.1190
cut4                    -1.0132         1.1107
scale                    1.9470        0.21005
```

This model suggests that respiratory response varies with `drug` and `base.` The negative parameter estimate for `trend` is not significant.

```
Log likelihood =     -697.88118     on     544 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
drug                  -0.70110         0.48780
male                  -0.28725         0.55095
age                   -0.16673E-01     0.16114E-01
base                   0.32507         0.10712
trend                  0.53462         0.20229
trend_drug            -0.38516         0.12083
cut1                   -5.4385          1.2199
cut2                   -3.8906          1.2006
cut3                   -1.5984          1.1850
cut4                   0.14683          1.1805
scale                   1.9802         0.21261
```

This model suggests that respiratory response varies with `base, trend` has a significant positive effect (for those on the placebo), while there is linear decline of respiratory status with visit (`trend`) for those on the treatment (`drug`). The main effect of `drug` which is negative, is not significant in this model.

We also need to remember that this is a highly selective sample, in that individuals who do not have respiratory illness are excluded. If the random effects for respiratory illness are independent of the covariates for epilepsy in the population, then this type of selectivity on outcome will have induced a correlation between the random effects and the included covariates, This correlation has not been allowed for in the analysis and our model is misspecified, e.g. by producing bias in the covariate parameters. Including `base` as a covariate complicates things further, this arises from the inclusion of `base` as an explanatory covariate as `base` can be treated as an endogenous initial condition for the response process. Further discussion of this issue is covered elsewhere.

11.2 Batch Script: `respiratory.do`

```
log using respiratory_s.log, replace
set more off
use respiratory2
#delimit ;
sabre, data ij r center drug male age bl v1 v2 v3 v4 patient status r1 r2 r3
        r4 r5 bld trend base;
sabre ij r center drug male age bl v1 v2 v3 v4 patient status r1 r2 r3 r4 r5
      bld trend base, read;
#delimit cr
sabre, case patient
sabre, yvar status
sabre, ordered y
sabre, lfit
sabre, dis m
sabre, dis e
sabre, quad a
sabre, mass 20
sabre, fit
sabre, dis m
```

```
sabre, dis e
sabre, fit drug male age base
sabre, dis m
sabre, dis e
sabre, fit drug male age base trend
sabre, dis m
sabre, dis e
sabre, trans trend_drug trend * drug
sabre, fit drug male age base trend trend_drug
sabre, dis m
sabre, dis e
log close
clear
exit
```

12 Exercise L8. Poisson Model of Epileptic Seizures

12.1 Relevant Results from `epilep_s.log` and Discussion

Task 1. Estimate a Poisson model for the response number of epileptic seizures (`y`) with a constant but without any random effects.

Result/Discussion

```
Log likelihood =     -1643.8739     on     235 residual degrees of freedom

    Parameter              Estimate         Std. Err.
    ___________________________________________________
    cons                     2.1118         0.22646E-01
```

Task 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.

Result/Discussion

```
Log likelihood =     -701.05330     on     234 residual degrees of freedom

    Parameter              Estimate         Std. Err.
    ___________________________________________________
    cons                     1.6213         0.12807
    scale                   0.94582         0.96382E-01
```

This model has a chi-square improvement of -2(-1643.8739+701.05330)= 1885. 6 over the homogeneous model. The sampling distribution of this test statistic is not chi-square with 1 df. Under the null hypothesis `scale` has the value 0, and it can only take values >0 under the alternative. The correct p value for this test statistics is obtained by dividing the naive p value of 1885. 6 for 1 degree of freedom by 1/2, and so its clearly significant.

Task 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?

Result/Discussion

```
    Log likelihood =     -665.58007     on     229 residual degrees of freedom

    Parameter              Estimate         Std. Err.
    ___________________________________________________
    cons                     2.1145         0.21972
```

```
lbas                  0.88443         0.13123
treat                -0.93304         0.40083
lbas_trt              0.33826         0.20334
lage                  0.48424         0.34728
visit                -0.29362         0.10142
scale                 0.50282         0.58625E-01
```

This gave a chi-square improvement over the previous model of -2(-701.05330 + 665.58007)= 70.946 for 234-229= 5 df, which is significant at the 0.05 level. But judged by the various covariate parameter estimates and their standard errors, the following main effects are not significant: `lbas_trt` and `lage`. The `scale` parameter in the model with covariates is still very significant, its nearly 1/2 the previous value but with a much smaller standard error.

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

Result/Discussion

```
Log likelihood =     -665.29074     on     229 residual degrees of freedom

Parameter              Estimate        Std. Err.
___________________________________________________
cons                    2.1143          0.21972
lbas                    0.88443         0.13123
treat                  -0.93304         0.40083
lbas_trt                0.33826         0.20334
lage                    0.48424         0.34728
v4                     -0.16109         0.54576E-01
scale                   0.50282         0.58625E-01
```

There is very little difference between the likelihood of this model, and that of Task 3. In terms of fit there is not much to choose between them. Both models use 1 parameter estimate for the variation over time. The real difference is in the way the models parameterise the variation over time; `visit` is a linear trend, while `v4` is just a binary indicator for the 4th visit. The similarity in fit suggests that most of the nonstationarity in the response sequence occurs at the last visit. Is this an end effect (bias report) that occurs at the finish of a trial that patients are sad to leave? A data set with a longer seizure sequence is needed to establish what is happening,

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

Result/Discussion

This model, which is clustered by patient (`subj`) over time, suggests that patient's with a high baseline (`lbas`) or `age` have a higher seizure rate. The coefficient on `visit` or `v4` is negative, as is the main effect on `treat`, i.e. these effects reduce the seizure rate. The interaction between treatment and baseline (`lbas_trt`) is not significant. The model could be simplified by removing `lbas_trt` and `lage`.

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

Result/Discussion

We could add the interaction effect of `treat` with `visit` or (`v4`), to examine whether the impact of treatment wears off. We could also try an interaction of the baseline `lbas` with `treat`, to test whether the effectiveness of the treatment differs with the severity of the condition.

There is an interesting modeling issue in this exercise, this arises from the inclusion of `lbas` as an explanatory covariate as `lbas` can be treated as an endogenous initial condition for the response process. Further discussion of this issue is covered elsewhere.

We also need to remember that this is a highly selective sample, in that individuals who do not have epileptic seizures are excluded. If the random effects for epilepsy are independent of the covariates for epilepsy in the population, then this type of selectivity on outcome will have induced a correlation between the random effects and the included covariates. This correlation has not been allowed for in the analysis and our model is misspecified, e.g. by producing bias in the covariate parameters. Including `lbas` as a covariate complicates things further.

12.2 Batch Script: `epilep.do`

```
log using epilep_s.log, replace
set more off
use epilep
sabre, data subj y treat visit v4 lage lbas lbas_trt constant
sabre subj y treat visit v4 lage lbas lbas_trt constant, read
sabre, case subj
sabre, yvar y
sabre, family p
sabre, constant cons
sabre, lfit cons
sabre, dis m
sabre, dis e
sabre, quad a
sabre, mass 12
sabre, fit cons
sabre, dis m
sabre, dis e
sabre, fit lbas treat lbas_trt lage visit cons
sabre, dis m
sabre, dis e
sabre, fit lbas treat lbas_trt lage v4 cons
sabre, dis m
sabre, dis e
log close
clear
exit
```

13 Exercise L9. Bivariate Linear Model of Expiratory Flow Rates

13.1 Relevant Results from `pefr_s.log` and Discussion

13.1.1 Standard Wright Meter: data set `pefr.dta`

Task 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.

Result/Discussion

```
Log likelihood =     -180.57200     on      30 residual degrees of freedom

Parameter                Estimate         Std. Err.
---------------------------------------------------
cons                       450.35           27.759
occ2                      -4.9412           5.1115
sigma                      14.903           2.5558
scale                      113.48           19.630
```

The 95% or 99% normal confidence intervals on `scale` with a S.E. 19.630 do not include 0. Similarly the z statistic for the null hypothesis that `scale` is 0, takes the value 113.48/19.630= 5.780 9.which greatly exceeds the critical value for a direction predicted z test at the 95% or 99% levels.

13.1.2 Mini Wright Meter: data set `pefr.dta`

Task 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.

Result/Discussion

```
Log likelihood =     -184.48885     on      30 residual degrees of freedom

Parameter                Estimate         Std. Err.
---------------------------------------------------
cons                       452.47           26.406
occ2                       2.8824           6.7935
sigma                      19.806           3.3967
scale                      107.06           18.677
```

The 95% or 99% normal confidence intervals on `scale` with a S.E. 18.677 do not include 0. Similarly the z statistic for the null hypothesis that `scale` is 0, takes the value 107.06/18.677= 5.732 2.which greatly exceeds the critical value for a direction predicted z test at the 95% or 99% levels.

13.1.3 Joint Model: data set `wp-wm.dta`

Task 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?

Result/Discussion

```
Log likelihood =     -343.56561     on      59 residual degrees of freedom

    Parameter              Estimate         Std. Err.
    ___________________________________________________
    r1                      450.35           27.759
    r1_occ2                -4.9412           5.1115
    r2                      452.47           26.406
    r2_occ2                 2.8824           6.7935
    sigma1                  14.903           2.5558
    sigma2                  19.806           3.3967
    scale1                  113.48           19.630
    scale2                  107.06           18.676
    corr                   0.97163          0.17255E-01
```

The 95% or 99% normal confidences intervals on `corr` with a S.E. 0.17255E-01 include 1 but do not include 0. The z statistic for the null hypothesis that `corr` is 0, takes the value 0.97163/0.017255= 56.31, which is clearly significant. The value of the estimates and standard errors for `r1_occ2` and `r2_occ2` from the joint analysis are the same as those obtained in Tasks 1 and 2

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

Result/Discussion

The very high correlation suggests that the two flow meters are equally good at measuring peak expiratory flow rate. Some other criterion, such as relative cost of flow meters would have to be used to make a decision between them. However, this is a very small sample, and the analysis should really be repeated in different contexts with larger samples before a decision made.

13.2 Batch Script: `pefr.do`

```
log using pefr_s.log, replace
set more off
use pefr
sabre, data id occasion wp wm occ2
```

```
sabre id occasion wp wm occ2, read
sabre, case id
sabre, yvar wp
sabre, family g
sabre, constant cons
sabre, quad a
sabre, mass 12
sabre, scale 110
sabre, fit occ2 cons
sabre, dis m
sabre, dis e
sabre, yvar wm
sabre, scale 100
sabre, fit occ2 cons
sabre, dis m
sabre, dis e
clear
use wp-wm
sabre, data ij r id occasion pefr occ2 r1 r2
sabre ij r id occasion pefr occ2 r1 r2, read
sabre, case id
sabre, yvar pefr
sabre, model b
sabre, rvar r
sabre, family first=g second=g
sabre, constant first=r1 second=r2
sabre, trans r1_occ2 r1 * occ2
sabre, trans r2_occ2 r2 * occ2
sabre, quad a
sabre, mass first=12 second=12
sabre, sigma first=20 second=20
sabre, scale first=110 second=100
sabre, rho 0.9
sabre, fit r1_occ2 r1 r2_occ2 r2
sabre, dis m
sabre, dis e
log close
clear
exit
```

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

14.1 Relevant Results from `wage-unionpan_s.log` and Discussion

14.1.1 Univariate models

14.1.2 Wage equation: data `wagepan.dta`

Task 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.

Result/Discussion

```
Log likelihood =     -2193.2846     on    4350 residual degrees of freedom

    Parameter               Estimate         Std. Err.
    ___________________________________________________
    cons                  -0.10783          0.11195
    educ                   0.10124          0.90191E-02
    black                 -0.14414          0.48198E-01
    hisp                   0.20187E-01      0.43128E-01
    exper                  0.11225          0.82472E-02
    expersq               -0.40754E-02      0.59074E-03
    married                0.62362E-01      0.16792E-01
    union                  0.10674          0.17872E-01
    sigma                  0.35120          0.40230E-02
    scale                  0.33018          0.11478E-01
```

The 95% or 99% normal confidence intervals on `scale` with a S.E. 0.11478E-01 do not include 0. Similarly the z statistic for the null hypothesis that `scale` is 0, takes the value 0.33018/0.011478= 28.766 which greatly exceeds the critical value for a direction predicted z test at the 95% or 99% levels.

14.1.3 Trade union membership: data `wagepan.dta`

Task 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?

Result/Discussion

```
    Log likelihood =     -1654.9281     on    4348 residual degrees of freedom

    Parameter               Estimate         Std. Err.
```

```
--------------------------------------------------
cons                    -2.4347          1.2006
black                    1.8871         0.47315
hisp                     1.1052         0.44739
exper                   -0.40595E-01    0.24199E-01
educ                    -0.60500E-01    0.93060E-01
poorhlth                -0.75608        0.50335
married                  0.34500        0.15984
rur                      0.20794        0.24023
nrthcen                  0.69825        0.38780
nrtheast                 0.87514        0.42444
south                    0.31154E-01    0.38514
scale                    3.0130         0.17885
```

The 95% or 99% normal confidence intervals on `scale` with a S.E. 0.17885 do not include 0. Similarly the z statistic for the null hypothesis that `scale` is 0, takes the value 3.0130/0.17885= 16. 847 which greatly exceeds the critical value for a direction predicted z test at the 95% or 99% levels.

14.1.4 Joint model: data `wage-unionpan.dta`

Task 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 12 for the linear model and mass 64 for the binary response model.

Result/Discussion

```
Log likelihood =     -3844.4397     on    8697 residual degrees of freedom

Parameter               Estimate        Std. Err.
--------------------------------------------------
r1                      -0.10219        0.11223
r1_educ                  0.10126        0.90413E-02
r1_black                -0.14102        0.48334E-01
r1_hisp                  0.21318E-01    0.43241E-01
r1_exper                 0.11179        0.82461E-02
r1_expersq              -0.40491E-02    0.59057E-03
r1_married               0.62457E-01    0.16778E-01
r1_union                 0.86886E-01    0.19234E-01
r2                      -2.5927          1.1917
r2_black                 1.8804         0.47009
r2_hisp                  1.1430         0.44495
r2_exper                -0.38736E-01    0.24185E-01
r2_educ                 -0.50835E-01    0.92232E-01
r2_poorhlth             -0.74877        0.50277
r2_married               0.32735        0.15948
r2_rur                   0.27268        0.24120
r2_nrthcen               0.75647        0.38587
```

```
r2_nrtheast              0.83701           0.42036
r2_south                 0.11396           0.38250
sigma1                   0.35112           0.40208E-02
scale1                   0.33116           0.11517E-01
scale2                    2.9962           0.17732
corr                     0.16309           0.58340E-01
```

Task 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?

Result/Discussion

The 95% or 99% normal confidences intervals on `corr` with a S.E. 0.58340E-01 do not include 0. The z statistic for the null hypothesis that `corr` is 0, takes the value 0.16309/0.058340= 2.795, which is significant at the 95% level. The estimated value of `corr` is 0.16309, implying a positive correlation between the random effects for log wages and trade union membership.

The parameter estimate on `union` in the log wage equation of Task 1 was 0.10674 (S.E. 0.17872E-01). In the joint model of Task 3 this becomes 0.86886E-01 (S.E. 0.19234E-01), i.e. smaller. Some of the magnitude of the estimated `union` parameter in the independent model of Task 1 has been taken up by the positive correlation of the random effects of the two response sequences in the joint model of Task 3. A larger `corr` would have had more impact. Had `corr` been negative, the estimate of the `union` effect in the wage equation of the joint model would have been bigger. There have been other minor changes, but nothing that is worthy of note.

14.2 Batch Script: `wage-unionpan.do`

```
log using wage-unionpan_s.log, replace
set more off
use wagepan
#delimit ;
sabre, data nr year agric black bus construc ent exper fin hisp poorhlth
            hours manuf married min nrthcen nrtheast occ1 occ2 occ3 occ4
            occ5 occ6 occ7 occ8 occ9 per pro pub rur south educ tra trad
            union lwage d81 d82 d83 d84 d85 d86 d87 expersq;
sabre nr year agric black bus construc ent exper fin hisp poorhlth hours
      manuf married min nrthcen nrtheast occ1 occ2 occ3 occ4 occ5 occ6 occ7
      occ8 occ9 per pro pub rur south educ tra trad union lwage d81 d82 d83
      d84 d85 d86 d87 expersq, read;
#delimit cr
sabre, case nr
sabre, yvar lwage
sabre, family g
sabre, constant cons
sabre, quad a
sabre, mass 12
sabre, fit educ black hisp exper expersq married union cons
sabre, dis m
sabre, dis e
```

```
sabre, yvar union
sabre, family b
sabre, mass 64
#delimit ;
sabre, fit black hisp exper educ poorhlth married rur nrthcen nrtheast south
          cons;
#delimit cr
sabre, dis m
sabre, dis e
clear
use wage-unionpan
#delimit ;
sabre, data ij r nr year agric black bus construc ent exper fin hisp
           poorhlth hours manuf married min nrthcen nrtheast occ1 occ2 occ3
           occ4 occ5 occ6 occ7 occ8 occ9 per pro pub rur south educ tra
           trad union lwage d81 d82 d83 d84 d85 d86 d87 expersq y r1 r2;
sabre ij r nr year agric black bus construc ent exper fin hisp poorhlth
      hours manuf married min nrthcen nrtheast occ1 occ2 occ3 occ4 occ5 occ6
      occ7 occ8 occ9 per pro pub rur south educ tra trad union lwage d81 d82
      d83 d84 d85 d86 d87 expersq y r1 r2;
#delimit cr
sabre, case nr
sabre, yvar y
sabre, model b
sabre, rvar r
sabre, family first=g
sabre, constant first=r1 second=r2
sabre, trans r1_educ r1 * educ
sabre, trans r1_black r1 * black
sabre, trans r1_hisp r1 * hisp
sabre, trans r1_exper r1 * exper
sabre, trans r1_expersq r1 * expersq
sabre, trans r1_married r1 * married
sabre, trans r1_union r1 * union
sabre, trans r2_black r2 * black
sabre, trans r2_hisp r2 * hisp
sabre, trans r2_exper r2 * exper
sabre, trans r2_educ r2 * educ
sabre, trans r2_poorhlth r2 * poorhlth
sabre, trans r2_married r2 * married
sabre, trans r2_rur r2 * rur
sabre, trans r2_nrthcen r2 * nrthcen
sabre, trans r2_nrtheast r2 * nrtheast
sabre, trans r2_south r2 * south
sabre, quad a
sabre, mass first=12 second=64
sabre, nvar 8
#delimit ;
sabre, fit r1_educ r1_black r1_hisp r1_exper r1_expersq r1_married r1_union
          r1
          r2_black r2_hisp r2_exper r2_educ r2_poorhlth r2_married r2_rur
          r2_nrthcen r2_nrtheast r2_south r2;
#delimit cr
sabre, dis m
sabre, dis e
log close
clear
exit
```

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

15.1 Relevant Results from `angina_s.log` and Discussion

Task 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. There is no point putting `dose` in the linear predictor for the model of pre-treatment data.

Result/Discussion

```
Log likelihood =     -347.61120     on   20981 residual degrees of freedom

Parameter              Estimate        Std. Err.
---------------------------------------------------
cons                    -10.365          1.1652
logt                    0.94104         0.21372
t2_dose                 -3.1632         0.98709
t3_dose                 -1.9604         0.88064
```

Task 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. Can the model be simplified? What does this result tell you?

Result/Discussion

```
Log likelihood =     -345.08870     on   20977 residual degrees of freedom

Parameter              Estimate        Std. Err.
---------------------------------------------------
cons                    -12.770          2.1821
t1_logt                  1.4132         0.39951
t2                       3.1187          3.1588
t2_logt                 0.61208         0.36732
t2_dose                -0.11826          2.2308
t3                       1.9366          3.1959
t3_logt                 0.97444         0.39177
t3_dose                 -1.2289          2.0951
```

This gave a chi-square improvement over the previous model of -2(-347.61120 + 345.08870)= 5.045 for 4 df, which is not significant at the 0.05 level. But judged by the various covariate parameter estimates and their standard errors, the following effects are not significant: `t2, t2_logt, t2_dose t3, t3_dose.` The only effects that are significant are `t1_logt`, `t3_logt.`

Task 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 questions 1 and 2?

Result/Discussion

```
Log likelihood =     -319.69936     on   20976 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
cons                   -37.671           6.3256
t1_logt                 6.1198           1.1582
t2                      16.820           5.6102
t2_logt                 3.0605          0.71798
t2_dose                -6.7730           4.3705
t3                      10.646           4.9201
t3_logt                 4.3845          0.89413
t3_dose                -7.5816           3.8103
scale                   2.8539          0.63481
```

Task 4. What is your preferred model and why?

Result/Discussion

The model of Task 3 is to be preferred over that of Task 2. The Task 3 model has a chi-square improvement of -2(-347.61120+319.69936)= 55.824 over the homogeneous model. The sampling distribution of this test statistic is not chi-square with 1 df. Under the null hypothesis `scale` has the value 0, and it can only take values >0 under the alternative. The correct p value for this test statistics is obtained by dividing the naive p value of 55.824 for 1 degree of freedom by 1/2, and so its clearly significant.

Relative to the model of Task 2, the pattern of significance in the covariate effects has changed, now the only effect that is not significant at the 95% level is `t2_dose.` The model of Task 3 also suggests that the higher the `dose,` the lower the probability of angina pectoris in an interval, even though its not significant at time 2. Perhaps `dose` takes more than 1 hour to be fully effective. The parameter estimates on `logt` suggest an increasing failure rate, i.e. the more intervals that have passed without angina pectoris, the more likely it is to happen.

A complication in interpreting all the results is the slight negative correlation between the initial response and `dose,` i.e. those subjects with shorter initial times to angina pectoris have been given larger doses.

15.2 Batch Script: `angina.do`

```
log using angina_s.log, replace
set more off
use angina
sabre, data id d time dose t y censored d1 d2 t1 t2 t3
sabre id d time dose t y censored d1 d2 t1 t2 t3, read
sabre, case id
sabre, yvar y
sabre, link c
sabre, constant cons
sabre, trans logt log t
sabre, trans t1_logt t1 * logt
sabre, trans t2_logt t2 * logt
sabre, trans t3_logt t3 * logt
sabre, trans t2_dose t2 * dose
sabre, trans t3_dose t3 * dose
sabre, lfit logt t2_dose t3_dose cons
sabre, dis m
sabre, dis e
sabre, lfit t1_logt t2 t2_logt t2_dose t3 t3_logt t3_dose cons
sabre, dis m
sabre, dis e
sabre, quad a
sabre, mass 24
sabre, fit t1_logt t2 t2_logt t2_dose t3 t3_logt t3_dose cons
sabre, dis m
sabre, dis e
log close
clear
exit
```

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

16.1 Relevant Results from `unemployed_s.log` and Discussion

Task 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, age2, age3, training`, `university.`

Result/Discussion

```
Log likelihood =     -863.34908     on    6054 residual degrees of freedom

Parameter              Estimate        Std. Err.
___________________________________________________
r1                     -0.65484        0.45936
r1_logt                -0.40989        0.83365E-01
r1_nation               0.10020        0.18813
r1_gender              -0.95154        0.17211
r1_age2                 0.29558        0.18359
r1_age3                -1.1159         0.28392
r1_training            -0.57196        0.17156
r1_uni                  0.39942        0.25236
r2                     -4.6425         0.87518
r2_logt                 0.71448E-01    0.16142
r2_nation              -1.3664         0.53701
r2_gender               0.27443        0.29517
r2_age2                -0.41115        0.43252
r2_age3                -2.8920          1.0148
r2_training            -0.90111E-01    0.33052
r2_uni                   1.7091        0.37030
```

There are quite a few significant effects in this model, for full time job there is: `r1_logt, r1_gender, r1_age3, r1_training`, and for part time job there is: `r2_nation, r2_age3, r2_uni.`

Task 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?

Result/Discussion

```
Log likelihood =     -858.28512     on    6052 residual degrees of freedom

Parameter              Estimate        Std. Err.
___________________________________________________
```

```
    r1                      -0.77929          0.54531
    r1_logt                 -0.25932          0.13074
    r1_nation                0.16157E-01      0.23254
    r1_gender                -1.0365          0.20469
    r1_age2                  0.35790          0.21942
    r1_age3                  -1.2412          0.32407
    r1_training             -0.63586          0.20499
    r1_uni                   0.54050          0.30442
    r2                       -6.4812           1.7150
    r2_logt                  0.47311          0.30686
    r2_nation                -2.0969          0.85421
    r2_gender                0.42721          0.42568
    r2_age2                 -0.49077          0.56016
    r2_age3                  -3.7307           1.3156
    r2_training              0.16193          0.45701
    r2_uni                    2.2742          0.67070
    scale1                   0.68982          0.26742
    scale2                    1.6341          0.57926
```

The model of Task 2 is to be preferred over that of Task 1. The Task 2 model has a chi-square improvement of -2(-863.34908+858.28512)= 10.128 over the homogeneous model. The sampling distribution of this test statistic is not chi-square with 2 df. Under the null hypothesis the two `scales` have the value 0, and they can only take values >0 under the alternative. The correct p value for this test statistics is obtained by dividing the naive p value of 10.128 for 2 degrees of freedom by 1/2, and so the `scale` effects are clearly significant.

Relative to the model of Task 1, the pattern of significance for the duration effects (`logt`) effects has changed. For transitions to full time job, `r1_logt` now has border line significance, `r2_logt` remains non significant. The covariates that were significant for Task 1 are still significant, i.e.: `r1_gender, r1_age3, r1_training`, and `r2_nation, r2_age3, r2_uni.`

Task 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?

Result/Discussion

```
Log likelihood =     -854.82180     on    6051 residual degrees of freedom

    Parameter               Estimate         Std. Err.
    ___________________________________________________
    r1                      -0.85561          0.55468
    r1_logt                 -0.26861          0.12096
    r1_nation                0.53793E-01      0.23762
    r1_gender                -1.0380          0.20881
    r1_age2                  0.37800          0.22498
    r1_age3                  -1.2128          0.32617
    r1_training             -0.65213          0.21040
    r1_uni                   0.54125          0.30705
```

```
r2                        -6.9983           1.8612
r2_logt                    0.34010          0.28645
r2_nation                 -2.2709           0.92092
r2_gender                  0.58557          0.45707
r2_age2                   -0.48868          0.57597
r2_age3                   -3.6769           1.3490
r2_training                0.30118          0.48255
r2_uni                     2.3040           0.71016
scale1                     0.78025          0.24496
scale2                     1.8157           0.59038
corr                      -1.0000           0.0000
```

Task 4. What is your preferred model and why?

Result/Discussion

We cant put 95% or 99% normal confidences intervals on `corr` as its S.E. is too small to be printed. However, the Task 3 model has a chi-square improvement of -2(-858.28512+854.82180)= 6.926 6 for 1 df over the independent model of Task 2, which is significant.

In the correlated model the pattern of significance has changed slightly. For transitions to full time job, `r1_logt` has become significant, `r2_logt` remains non significant. The covariates that were significant for Task 2 are still significant, i.e.: `r1_gender, r1_age3, r1_training`, and `r2_nation, r2_age3, r2_uni`. In both transitions age3 has a large negative values, suggesting that the older unemployed are less likely to find employment of any kind. The large negative correlation in the random effects is a manifestation of single spell competing risk data, i.e. if a transition from unemployment to full time job occurs, then the transition to part time job cannot occur.

This analysis also ignores a selection problem that occurs with an analysis that is restricted to specifc flows, i.e. does not simultaneously consider all the transitions, e.g. from the origin, part time work. If the random effects and observed covariates for labour behaviour are independent in the population, then the random effects and observed covariates for any specific flow or subset of flows will be correlated, see Chesher A. & Lancaster T., (1981), Stock and Flow Sampling, Economics Letters, Vol. 8, 63-65, for further details. As this correlation is not taken into account by the model, the parameter estimates will biased. A complement of this problem occurs if the random effects and observed covariates are correlated in the population, then they could be either less or even more correlated in specific flows. Consequently, its probably best to compare inferences from both the joint and separate analysis of all the flows with the proposed state space.

16.2 Batch Script: `unemployed.do`

```
log using unemployed_s.log, replace
set more off
use unemployed
#delimit ;
sabre, data id t survival full part nationality gender age training
           university rowname spell y r r1 r2 id_spell age1 age2 age3;
```

```
sabre id t survival full part nationality gender age training university
      rowname spell y r r1 r2 id_spell age1 age2 age3, read;
#delimit cr
sabre, case id
sabre, yvar y
sabre, model b
sabre, rvar r
sabre, link first=c second=c
sabre, constant first=r1 second=r2
sabre, trans logt log t
sabre, trans r1_logt r1 * logt
sabre, trans r1_nation r1 * nationality
sabre, trans r1_gender r1 * gender
sabre, trans r1_age2 r1 * age2
sabre, trans r1_age3 r1 * age3
sabre, trans r1_training r1 * training
sabre, trans r1_uni r1 * university
sabre, trans r2_logt r2 * logt
sabre, trans r2_nation r2 * nationality
sabre, trans r2_gender r2 * gender
sabre, trans r2_age2 r2 * age2
sabre, trans r2_age3 r2 * age3
sabre, trans r2_training r2 * training
sabre, trans r2_uni r2 * university
sabre, nvar 8
#delimit ;
sabre, lfit r1 r1_logt r1_nation r1_gender r1_age2 r1_age3 r1_training
            r1_uni
            r2 r2_logt r2_nation r2_gender r2_age2 r2_age3 r2_training
            r2_uni;
#delimit cr
sabre, dis m
sabre, dis e
sabre, quad a
sabre, mass first=32 second=32
sabre, corr n
sabre, nvar 8
#delimit ;
sabre, fit r1 r1_logt r1_nation r1_gender r1_age2 r1_age3 r1_training r1_uni
           r2 r2_logt r2_nation r2_gender r2_age2 r2_age3 r2_training r2_uni
           ;
#delimit cr
sabre, dis m
sabre, dis e
sabre, corr y
sabre, nvar 8
#delimit ;
sabre, fit r1 r1_logt r1_nation r1_gender r1_age2 r1_age3 r1_training r1_uni
           r2 r2_logt r2_nation r2_gender r2_age2 r2_age3 r2_training r2_uni
           ;
#delimit cr
sabre, dis m
sabre, dis e
log close
clear
exit
```

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

17.1 Relevant Results from `manager_s.log` and Discussion

Task 1. Estimate a linear model (without random effects) for the `scores` with the pupil- and school- level covariates `dirsex`, `schtype` and `pupsex`.

Result/Discussion

```
Log likelihood =     -7758.0889     on    4975 residual degrees of freedom

Parameter              Estimate         Std. Err.
---------------------------------------------------
cons                     2.1708         0.70508E-01
dirsex                  0.91255E-01     0.32600E-01
fschtype    (    1)      0.0000         ALIASED [I]
fschtype    (    2)     0.37444         0.38193E-01
fschtype    (    3)     0.15259         0.43772E-01
pupsex                 -0.21601E-01     0.33829E-01
sigma                    1.1492
```

The covariate `fschtype` is the factor variable for `schtype, fschtype(1)` is `ALIASED` because the model contains a constant.

Task 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?

Result/Discussion

```
Log likelihood =     -7272.8266     on    4974 residual degrees of freedom

Parameter              Estimate         Std. Err.
---------------------------------------------------
cons                     2.1638         0.11778
dirsex                  0.10048         0.54458E-01
fschtype    (    1)      0.0000         ALIASED [I]
fschtype    (    2)     0.39401         0.63790E-01
fschtype    (    3)     0.19282         0.72611E-01
pupsex                 -0.21618E-01     0.56559E-01
sigma                   0.91863         0.10132E-01
scale                   0.69752         0.22281E-01
```

The log likelihood of the homogeneous model of Task 1 is -7758.0889, and log likelihood of the random effects model of Task 2 is -7272.8266. The change in log likelihood over the homogeneous model is -2(-7758.0889+7272.8266)= 970.

52. The sampling distribution of this test statistic is not chi-square with 1 df. Under the null hypothesis `scale` has the value 0, and it can only take values >0 under the alternative. The correct p value for this test statistics is obtained by dividing the naive p value of 970.52 for 1 degree of freedom by 1/2, and so its clearly significant, suggesting that the `scores` from pupils to 6 different questions are highly correlated.

Task 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?

Result/Discussion

```
Log likelihood =     -7223.1596     on    4973 residual degrees of freedom

    Parameter              Estimate         Std. Err.
    ___________________________________________________
    cons                     2.2429         0.16818
    dirsex                  0.10251         0.92085E-01
    fschtype    (    1)      0.0000         ALIASED [I]
    fschtype    (    2)     0.39067         0.10834
    fschtype    (    3)     0.19933         0.12026
    pupsex                -0.77852E-01      0.53255E-01
    sigma                   0.91881         0.10137E-01
    scale2                  0.58396         0.21798E-01
    scale3                  0.38029         0.38309E-01
```

The log likelihood of the homogeneous model of Task 1 is -7758.0889, and the log likelihood of the 3-level random effects model of Task 3 is -7223.1596. The change in log likelihood over the homogeneous model is -2(-7758.0889 + 7223.1596)= 1069.9. The sampling distribution of this test statistic is not chi-square with 2 df. The null hypothesis is that `scale2` and `scale3` have the value 0, they can only take values >0 under the alternative. The correct p value for this test statistics is obtained by dividing the naive p value of 1069.9 for 2 degrees of freedom by 1/2, and so its clearly significant, suggesting that the `scores` from pupils to 6 different questions with the same school are highly correlated. The highest correlation occurs between `scores` of the same pupil than between `scores` of different pupils in the same school, as `scale2` is greater than `scale3`

The log likelihood of the 2-level model of Task 2 is -7272.8266, and log likelihood of the 3-level model of Task 3 is -7223.1596. The change in log likelihood over the Task 2 model is -2(-7272.8266+7223.1596)= 99.334. The sampling distribution of this test statistic is not chi-square with 1 df. Under the null hypothesis that `scale3` have the value 0, and it can only take values >0 under the alternative. The correct p value for this test statistics is obtained by dividing the naive p value of 99.334 for 1 degrees of freedom by 1/2, and so its clearly significant.

Task 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

Result/Discussion?

The significant covariates in the Task 1 and 2 models are: `fschtype(2)`, `fschtype(3)`, but only `fschtype(2)` remains significant in the Task 3 model. The main change as we move from the Task 1 to the Task 2 model, is that the standard errors of the covariates become noticeably larger. The standard errors tended to become larger again as we moved from the Task2 to the Task 3 results.

17.2 Batch Script: `manager.do`

```
log using manager_s.log, replace
set more off
use manager
sort id
sabre, data id school pupil dirsex schtype pupsex item constant class scores
sabre id school pupil dirsex schtype pupsex item constant class scores, read
sabre, case id
sabre, yvar scores
sabre, family g
sabre, constant cons
sabre, fac schtype fschtype
sabre, lfit dirsex fschtype pupsex cons
sabre, dis m
sabre, dis e
sabre, quad a
sabre, mass 12
sabre, fit dirsex fschtype pupsex cons
sabre, dis m
sabre, dis e
clear
use manager
sabre, data id school pupil dirsex schtype pupsex item constant class scores
sabre id school pupil dirsex schtype pupsex item constant class scores, read
sabre, case first=id second=school
sabre, yvar scores
sabre, family g
sabre, constant cons
sabre, fac schtype fschtype
sabre, quad a
sabre, mass first=24 second=24
sabre, fit dirsex fschtype pupsex cons
sabre, dis m
sabre, dis e
log close
clear
exit
```

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

18.1 Relevant Results from `towerl_s.log` and Discussion

Task 1. Estimate a logit model (without random effects, use `lfit`) for the binary response `dtlm` with the covariate `level`, and dummy variables for `group`=2 and `group`=3.

Result/Discussion

```
Log likelihood =     -313.89079     on     673 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
cons                    -1.1605         0.18245
level                   -1.3134         0.14095
fgroup      (   1)       0.0000         ALIASED [I]
fgroup      (   2)    -0.13966          0.22825
fgroup      (   3)    -0.83133          0.27423
```

The covariate `fgroup` is the factor variable for `group, fgroup(1)` is `ALIASED` because the model contains a constant.

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

Result/Discussion

```
Log likelihood =     -305.95929     on     672 residual degrees of freedom

    Parameter              Estimate         Std. Err.
    ___________________________________________________
    cons                    -1.4827         0.28356
    level                   -1.6488         0.19335
    fgroup      (   1)       0.0000         ALIASED [I]
    fgroup      (   2)    -0.16907          0.33425
    fgroup      (   3)     -1.0227          0.39385
    scale                    1.2943         0.25571
```

The log likelihood of the homogeneous model of Task 1 is -313.89079, and log likelihood of the random effects model of Task 2 is -305.95929. The change in log likelihood over the homogeneous model is -2(-313.89079 + 305.95929)= 15.863. The sampling distribution of this test statistic is not chi-square with 1 df. Under the null hypothesis `scale` has the value 0, it can only take values >0 under the alternative. The correct p value for this test statistics is obtained by dividing the naive p value of 15.863 for 1 degree of freedom by 1/2, and so its clearly significant, suggesting that the `dtlm` values from subjects at 3 different occasions are highly correlated.

Task 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?

Result/Discussion

```
Log likelihood =     -305.12036     on     671 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
cons                    -1.4859         0.28486
level                   -1.6485         0.19322
fgroup      (    1)      0.0000         ALIASED [I]
fgroup      (    2)    -0.24867         0.35440
fgroup      (    3)     -1.0523         0.39999
scale2                   1.0668         0.32154
scale3                  0.75445         0.34591
```

The log likelihood of the homogeneous model of Task 1 is -313.89079, and the log likelihood of the 3-level random effects model of Task 3 is -305.12036. The change in log likelihood over the homogeneous model is -2(-313.89079 +305.12036)= 17.541 The sampling distribution of this test statistic is not chi-square with 2 df. The null hypothesis is that `scale2` and `scale3` have the value 0, they can only take values >0 under the alternative. The correct p value for this test statistics is obtained by dividing the naive p value of 17.541 for 2 degrees of freedom by 1/2, and so its clearly significant, suggesting that the `dtlm` values from subjects at 3 different occasions with the same family are correlated.

The log likelihood of the 2-level model of Task 2 is -305.95929, and log likelihood of the 3-level model of Task 3 is -305.12036. The change in log likelihood over the Task 2 model is -2(-305.95929+305.12036)= 1.6779. The sampling distribution of this test statistic is not chi-square with 1 df. Under the null hypothesis that `scale3` has the value 0, it can only take values >0 under the alternative. The correct p value for this test statistics is obtained by dividing the naive p value of 1.6779 for 1 degrees of freedom by 1/2, and so its not a significant improvement over the model of Task 2

Task 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?

Result/Discussion

The significant covariates in the Task 1, 2 and 3 models are: `level,` and `fgroup(3).` The main change as we move from the Task 1 to the Task 2 model, is that the estimates and standard errors become larger, this is one of the features of a binary response model with significant random effects. Even though a 95% confidence interval on scale3 does not include the value 0, we would take the likelihood ratio test for the model of Task2 against the model of Task 3 as a more reliable indicator of significance.

18.2 Batch Script: `towerl.do`

```
log using towerl_s.log, replace
set more off
use towerl
sort id
#delimit ;
sabre, data id level famnum group age sex tlm tlpl tlcpl tlsub tlcsub occ
            dtlm;
sabre id level famnum group age sex tlm tlpl tlcpl tlsub tlcsub occ dtlm,
      read;
#delimit cr
sabre, case id
sabre, yvar dtlm
sabre, constant cons
sabre, fac group fgroup
sabre, lfit level fgroup cons
sabre, dis m
sabre, dis e
sabre, quad a
sabre, mass 12
sabre, fit level fgroup cons
sabre, dis m
sabre, dis e
clear
use towerl
#delimit ;
sabre, data id level famnum group age sex tlm tlpl tlcpl tlsub tlcsub occ
            dtlm;
sabre id level famnum group age sex tlm tlpl tlcpl tlsub tlcsub occ dtlm,
      read;
#delimit cr
sabre, case first=id second=famnum
sabre, yvar dtlm
sabre, constant cons
sabre, fac group fgroup
sabre, mass first=12 second=12
sabre, fit level fgroup cons
sabre, dis m
sabre, dis e
log close
clear
exit
```

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

19.1 Relevant Results from `guatemala_immun_s.log` and Discussion

Task 1. Estimate a logit model (without random effects, use `lfit` with a constant for the binary response `immun` with the covariates `kid2p, mom25p, order23, order46, order7p, indnospa, indspa, momedpri, momedsec, husedpri, husedsec, huseddk, momwork, rural` and `pcind81`.

Result/Discussion

```
Log likelihood =     -1399.5897     on    2143 residual degrees of freedom

Parameter             Estimate         Std. Err.
___________________________________________________
cons                 -0.72573         0.21946
kid2p                 0.95096         0.11437
mom25p               -0.78252E-01     0.12141
order23              -0.83857E-01     0.13429
order46               0.92846E-01     0.15967
order7p               0.15486         0.19721
indnospa              0.27805         0.19899
indspa                0.21984         0.16372
momedpri              0.24986         0.10575
momedsec              0.29884         0.23791
husedpri              0.28872         0.10994
husedsec              0.21011         0.19872
huseddk               0.32750E-01     0.17710
momwork               0.24757         0.95179E-01
rural                -0.49695         0.11418
pcind81              -0.77611         0.20570
```

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

Result/Discussion

```
Log likelihood =     -1339.3508     on    2142 residual degrees of freedom

Parameter             Estimate         Std. Err.
___________________________________________________
cons                  -1.2768         0.43706
kid2p                  1.7261         0.21823
```

```
mom25p                  -0.21704          0.23276
order23                 -0.26755          0.23411
order46                  0.10310          0.29648
order7p                  0.35413          0.37359
indnospa                 0.48022          0.40812
indspa                   0.31757          0.33314
momedpri                 0.53171          0.22215
momedsec                 0.57291          0.48630
husedpri                 0.52739          0.22910
husedsec                 0.40611          0.41083
huseddk                 -0.68018E-02      0.36130
momwork                  0.47754          0.19918
rural                   -0.91104          0.24219
pcind81                 -1.3932           0.42842
scale                    2.5036           0.27063
```

The log likelihood of the homogeneous model of Task 1 is -1399.5897, and log likelihood of the random effects model of Task 2 is -1339.3508. The change in log likelihood over the homogeneous model is -2(-1399.5897 + 1339.3508)= 120. 48. The sampling distribution of this test statistic is not chi-square with 1 df. Under the null hypothesis `scale` has the value 0, it can only take values >0 under the alternative. The correct p value for this test statistics is obtained by dividing the naive p value of 120. 48 for 1 degree of freedom by 1/2, and so its clearly significant, suggesting that the `immun` values from `kid`s from the same family (`mom`) are highly correlated.

Task 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?

Result/Discussion

```
Log likelihood =     -1323.9524     on    2141 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
cons                    -1.2362           0.48246
kid2p                    1.7174           0.21750
mom25p                  -0.21457          0.23155
order23                 -0.26133          0.23197
order46                  0.17784          0.29446
order7p                  0.43080          0.37227
indnospa                -0.17518          0.48971
indspa                  -0.83921E-01      0.36352
momedpri                 0.43242          0.22239
momedsec                 0.41924          0.48397
husedpri                 0.54095          0.23248
husedsec                 0.50729          0.41425
```

```
huseddk              -0.60728E-02       0.35689
momwork               0.39027           0.20279
rural                -0.88619           0.30507
pcind81              -1.1512            0.50069
scale2                2.3172            0.26215
scale3                1.0249            0.15995
```

The log likelihood of the homogeneous model of Task 1 is -1399.5897, and the log likelihood of the 3-level random effects model of Task 3 is -1323.9524. The change in log likelihood over the homogeneous model is -2(-1399.5897 +1323.9524)= 151.27 The sampling distribution of this test statistic is not chi-square with 2 df. The null hypothesis is that `scale2` and `scale3` have the value 0, they can only take values >0 under the alternative. The correct p value for this test statistics is obtained by dividing the naive p value of 151.27 for 2 degrees of freedom by 1/2, and so its clearly significant, suggesting that the `immun` values from kids in the same family and from different families in the same community are correlated.

The log likelihood of the 2-level model of Task 2 is -305.95929, and log likelihood of the 3-level model of Task 3 is -305.12036. The change in log likelihood over the Task 2 model is -2(-1339.3508+1323.9524)= 30.797. The sampling distribution of this test statistic is not chi-square with 1 df. Under the null hypothesis that `scale3` has the value 0, it can only take values >0 under the alternative. The correct p value for this test statistics is obtained by dividing the naive p value of 30.797 for 1 degrees of freedom by 1/2, and so its a significant improvement over the model of Task 2

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

Result/Discussion

The same covariates: `kid2p, momedpri, husedpri, momwork, rural,` and `pcind81` are more or less significant in all 3 models, the main difference is that in the Task 3 model, `momedpri` and `momwork` are marginal. The main change as we move on from the Task 1, Task 2 and Task 3 models, is that there is a tendency for estimates and standard errors become larger, this is one of the features of a binary response model with significant random effects. Though this effect is not allways that clear between Task 2 and 3, for instance the parameter estimate on `kid2p` from the model of Task 1 is 0.95096 (S.E.0.11437), Task 2 is 1.7261 (S.E. 0.21823), while that from the model of Task 3 is 1.7174 (S.E. 0.21750)

19.2 Batch Script: `guatemala_immun.do`

```
log using guatemala_immun_s.log, replace
set more off
use guatemala_immun
#delimit ;
sabre, data kid mom cluster immun kid2p mom25p order23 order46 order7p
          indnospa indspa momedpri momedsec husedpri husedsec huseddk
          momwork rural pcind81;
```

```
sabre kid mom cluster immun kid2p mom25p order23 order46 order7p indnospa
      indspa momedpri momedsec husedpri husedsec huseddk momwork rural
      pcind81, read;
#delimit cr
sabre, case mom
sabre, yvar immun
sabre, constant cons
#delimit ;
sabre, lfit kid2p mom25p order23 order46 order7p indnospa indspa momedpri
            momedsec husedpri husedsec huseddk momwork rural pcind81 cons;
#delimit cr
sabre, dis m
sabre, dis e
sabre, quad a
sabre, mass 24
#delimit ;
sabre, fit kid2p mom25p order23 order46 order7p indnospa indspa momedpri
           momedsec husedpri husedsec huseddk momwork rural pcind81 cons;
#delimit cr
sabre, dis m
sabre, dis e
sabre, case first=mom second=cluster
sabre, quad a
sabre, mass first=32 second=32
#delimit ;
sabre, fit kid2p mom25p order23 order46 order7p indnospa indspa momedpri
           momedsec husedpri husedsec huseddk momwork rural pcind81 cons;
#delimit cr
sabre, dis m
sabre, dis e
log close
clear
exit
```

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

20.1 Relevant Results from `deaths_s.log` and Discussion

Task 1. Estimate a Poisson model (without random effects, use `lfit`) for the number of deaths (`deaths`) with the covariate `uvb`. Use log `expected` deaths as an offset.

Result/Discussion

```
Log likelihood =     -1723.7727     on     351 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
cons                 -0.70104E-01      0.11047E-01
uvb                  -0.57191E-01      0.26770E-02
```

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

Result/Discussion

```
Log likelihood =     -1125.1505     on     351 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
cons                 -0.13860          0.49393E-01
uvb                  -0.34415E-01      0.10038E-01
scale                 0.41217          0.37598E-01
```

The log likelihood of the homogeneous model of Task 1 is −1723.7727, and log likelihood of the random effects model of Task 2 is -1339.3508. The change in log likelihood over the homogeneous model is -2(-1723.7727 + 1125.1505)= 1197. 2. The sampling distribution of this test statistic is not chi-square with 1 df. Under the null hypothesis `scale` has the value 0, it can only take values >0 under the alternative. The correct p value for this test statistics is obtained by dividing the naive p value of 1197. 2 for 1 degree of freedom by 1/2, and so its clearly significant, suggesting that the `death` values from different `counties` from the same family (`region`) are highly correlated.

Task 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?

Result/Discussion

```
Log likelihood =     -1095.3100     on     350 residual degrees of freedom

    Parameter              Estimate         Std. Err.
    ___________________________________________________
    cons                  -0.63968E-01      0.13358
    uvb                   -0.28204E-01      0.11400E-01
    scale2                 0.21988          0.24804E-01
    scale3                 0.37037          0.97658E-01
```

The log likelihood of the homogeneous model of Task 1 is -1723.7727, and the log likelihood of the 3-level random effects model of Task 3 is -1095.3100. The change in log likelihood over the homogeneous model is -2(-1723.7727 + 1095.3100)= 1256. 9 The sampling distribution of this test statistic is not chi-square with 2 df. The null hypothesis is that `scale2` and `scale3` have the value 0, they can only take values >0 under the alternative. The correct p value for this test statistics is obtained by dividing the naive p value of 1256. 9 for 2 degrees of freedom by 1/2, and so its clearly significant, suggesting that the `death` values from different `counties` from the same family (`region`), and from different `regions` in the same `nation` are highly correlated.

The log likelihood of the 2-level model of Task 2 is -1125.1505, and log likelihood of the 3-level model of Task 3 is -1095.3100. The change in log likelihood over the Task 2 model is -2(-1125.1505+1095.3100)= 59. 681. The sampling distribution of this test statistic is not chi-square with 1 df. Under the null hypothesis that `scale3` has the value 0, it can only take values >0 under the alternative. The correct p value for this test statistics is obtained by dividing the naive p value of 59. 681 for 1 degrees of freedom by 1/2, and so its a significant improvement over the model of Task 2

Task 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?

Result/Discussion

The z statistics for `uvb` from the model of Task 1 is -0.057191/0.0026770= -21.364, Task 2 is -0.034415/0.010038= -3.4285, while that from the model of Task 3 is -0.028204/0.011400= -2.474, i.e. the estimates decline and become a lot less less significant (S.E.s increase) as higher level random effects are added.

20.2 Batch Script: `deaths.do`

```
log using deaths_s.log, replace
set more off
use deaths
sabre, data nation region county deaths expected uvb mr
sabre nation region county deaths expected uvb mr, read
sabre, case region
sabre, yvar deaths
sabre, family p
sabre, constant cons
sabre, trans logexp log expected
sabre, offset logexp
sabre, lfit uvb cons
```

```
sabre, dis m
sabre, dis e
sabre, quad a
sabre, mass 12
sabre, fit uvb cons
sabre, dis m
sabre, dis e
sabre, case first=region second=nation
sabre, quad a
sabre, mass first=96 second=96
sabre, fit uvb cons
sabre, dis m
sabre, dis e
log close
clear
exit
```

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

21.1 Relevant Results from `vwks_s.log` and Discussion

Task 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`).

Result/Discussion

```
 Log likelihood =     -2340.6156     on   28773 residual degrees of freedom

    Parameter              Estimate         Std. Err.
    ___________________________________________________
    ft          (   1)      -7.3253          0.76287
    ft          (   2)      -7.6077          0.76647
    ft          (   3)      -8.1945          0.76760
    ft          (   4)      -8.4380          0.77476
    ft          (   5)      -9.1986          0.80081
    ft          (   6)      -9.4309          0.78929
    ft          (   7)      -9.0874          0.77870
    ft          (   8)      -9.3464          0.79907
    ft          (   9)      -9.7955          0.83441
    ft          (  10)      -10.490          0.89070
    loguu                   0.74703          0.83416E-01
    logvv                  -0.15591          0.76683E-01
    nonman                 -0.19363          0.10924
    written                -0.67264          0.11567
    size                    0.27550E-01      0.36976E-01
    wage                   -0.24750E-01      0.51028E-01
    grade                   0.86721E-01      0.54348E-01
    dayrel                 -0.39327          0.12075
```

The covariate `ft(.)` is the factor variable for `t`, there is no constant in the model.

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

Result/Discussion

```
 Log likelihood =     -2268.2074     on   28772 residual degrees of freedom

    Parameter              Estimate         Std. Err.
```

```
--------------------------------------------------------
ft            (    1)     -10.660            1.3780
ft            (    2)     -10.458            1.3499
ft            (    3)     -10.728            1.3365
ft            (    4)     -10.715            1.3324
ft            (    5)     -11.294            1.3435
ft            (    6)     -11.318            1.3329
ft            (    7)     -10.756            1.3412
ft            (    8)     -10.643            1.3635
ft            (    9)     -10.883            1.3841
ft            (   10)     -11.280            1.4424
loguu                      1.0886           0.15437
logvv                     -0.26518          0.13096
nonman                    -0.44384          0.19154
written                   -0.94262          0.21713
size                       0.87120E-01      0.63396E-01
wage                       0.60059E-01      0.91802E-01
grade                      0.56564E-01      0.10113
dayrel                    -0.66028          0.22303
scale                       1.9924          0.20134
```

The log likelihood of the homogeneous model of Task 1 is -2340.6156, and log likelihood of the random effects model of Task 2 is -2268.2074. The change in log likelihood over the homogeneous model is -2(-2340.6156 + 2268.2074)= 144.82. The sampling distribution of this test statistic is not chi-square with 1 df. Under the null hypothesis `scale` has the value 0, it can only take values >0 under the alternative. The correct p value for this test statistics is obtained by dividing the naive p value of = 144.82 for 1 degree of freedom by 1/2, and so its clearly significant, suggesting that the binary response values (`match`) from different `weeks` from the same `vacancy` are highly correlated.

Task 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?

Result/Discussion

```
Log likelihood =      -2247.6656     on   28771 residual degrees of freedom

Parameter                Estimate         Std. Err.

--------------------------------------------------------
ft            (    1)     -9.7980            1.4117
ft            (    2)     -9.6039            1.3854
ft            (    3)     -9.8799            1.3725
ft            (    4)     -9.8826            1.3689
ft            (    5)     -10.452            1.3803
```

```
ft          (   6)     -10.451           1.3703
ft          (   7)     -9.8342           1.3806
ft          (   8)     -9.6961           1.4088
ft          (   9)     -9.8826           1.4293
ft          (  10)     -10.246           1.4852
loguu                   1.1429          0.16637
logvv                  -0.48556         0.14794
nonman                 -0.44829         0.20378
written                -0.79079         0.22718
size                    0.72855E-01     0.78514E-01
wage                    0.11520E-01     0.95085E-01
grade                   0.15733E-01     0.10515
dayrel                 -0.66339         0.23044
scale2                  1.5626          0.19974
scale3                  1.2265          0.15780
```

The log likelihood of the homogeneous model of Task 1 is -2340.6156, and the log likelihood of the 3-level random effects model of Task 3 is -2247.6656. The change in log likelihood over the homogeneous model is -2(-2340.6156 + 2247.6656)= 185.9 The sampling distribution of this test statistic is not chi-square with 2 df. The null hypothesis is that `scale2` and `scale3` have the value 0, they can only take values >0 under the alternative. The correct p value for this test statistics is obtained by dividing the naive p value of 185.9 for 2 degrees of freedom by 1/2, and so its clearly significant, suggesting that the the binary response values (`match`) from different `weeks` from the same `vacancy` are highly correlated and similarly from different `vacancies` of the same employer (`empref`) are highly correlated.

The log likelihood of the 2-level model of Task 2 is -2268.2074, and log likelihood of the 3-level model of Task 3 is -2247.6656. The change in log likelihood over the Task 2 model is -2(-2268.2074+2247.6656)= 41.084. The sampling distribution of this test statistic is not chi-square with 1 df. Under the null hypothesis that `scale3` has the value 0, it can only take values >0 under the alternative. The correct p value for this test statistics is obtained by dividing the naive p value of 41.084 for 1 degrees of freedom by 1/2, and so its a significant improvement over the model of Task 2.

Task 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?

Result/Discussion

The same external covariates are significant in all Tasks, namely: `loguu, logvv, nonman, written, dayrel.` The main change as we move from the Task 1 to the Task 2 model, is that both the magnitude of the estimate and the standard errors of the covariates become noticeably larger. The same happens again as we move from the Task 2 to the Task 3 results.

The dummy variables for vacancy duration `ft()` are also significant in all Tasks. The estimates on the various levels of vacancy duration also tend to increase in magnitude and their standard errors increase as we add more levels.

21.2 Batch Script: vwks.do

```
log using vwks_s.log, replace
set mem 100m
set more off
use vwks4_30k
#delimit ;
sabre, data notify order tape exposure elapsed1 elapsed2 ncon ncons match
            lad14 tape1 adjust wk week search lad skill nonman written size
            wage waged wages month u uu v vv vacref grade minage maxage
            office notified remain inter empref onthejob dayrel appren
            inhouse othertr notrain t t1 loguu logvv logu logv new old logu1
            loguu1 logv1 logvv1 logu2 loguu2 logv2 logvv2;
sabre notify order tape exposure elapsed1 elapsed2 ncon ncons match lad14
      tape1 adjust wk week search lad skill nonman written size wage waged
      wages month u uu v vv vacref grade minage maxage office notified
      remain inter empref onthejob dayrel appren inhouse othertr notrain t
      t1 loguu logvv logu logv new old logu1 loguu1 logv1 logvv1 logu2
      loguu2 logv2 logvv2, read;
#delimit cr
sabre, case vacref
sabre, yvar match
sabre, link c
sabre, fac t ft
sabre, lfit ft loguu logvv nonman written size wage grade dayrel
sabre, dis m
sabre, dis e
sabre, quad a
sabre, mass 48
sabre, fit ft loguu logvv nonman written size wage grade dayrel
sabre, dis m
sabre, dis e
sort empref vacref wk
#delimit ;
sabre, data notify order tape exposure elapsed1 elapsed2 ncon ncons match
            lad14 tape1 adjust wk week search lad skill nonman written size
            wage waged wages month u uu v vv vacref grade minage maxage
            office notified remain inter empref onthejob dayrel appren
            inhouse othertr notrain t t1 loguu logvv logu logv new old logu1
            loguu1 logv1 logvv1 logu2 loguu2 logv2 logvv2;
sabre notify order tape exposure elapsed1 elapsed2 ncon ncons match lad14
      tape1 adjust wk week search lad skill nonman written size wage waged
      wages month u uu v vv vacref grade minage maxage office notified
      remain inter empref onthejob dayrel appren inhouse othertr notrain t
      t1 loguu logvv logu logv new old logu1 loguu1 logv1 logvv1 logu2
      loguu2 logv2 logvv2, read;
#delimit cr
sabre, case first=vacref second=empref
sabre, yvar match
sabre, link c
sabre, fac t ft
sabre, quad a
sabre, mass first=64 second=64
sabre, fit ft loguu logvv nonman written size wage grade dayrel
sabre, dis m
sabre, dis e
log close
clear
exit
```

22 Exercise EP1. Trade Union Membership with Endpoints

22.1 Relevant Results from `nlsunion_end_s.log` and Discussion

Task 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.

Result/Discussion

```
Log likelihood =     -7641.6559     on   18986 residual degrees of freedom

  Parameter              Estimate         Std. Err.
  ---------------------------------------------------
  cons                    -2.6788         0.39094
  age                     0.22961E-01     0.23695E-01
  age2                   -0.22716E-03     0.37805E-03
  black                   0.84389         0.72350E-01
  msp                    -0.65237E-01     0.41003E-01
  grade                   0.70700E-01     0.12640E-01
  not_smsa               -0.11693         0.59975E-01
  south                  -0.74693         0.58813E-01
  scale                    1.5077         0.40779E-01
```

Task 2. Re-estimate 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?

Result/Discussion

```
Log likelihood =     -7632.6474     on   18985 residual degrees of freedom

Parameter              Estimate         Std. Err.
---------------------------------------------------
cons                    -2.7029         0.38943
age                     0.22211E-01     0.23671E-01
age2                   -0.21579E-03     0.37757E-03
black                   0.85198         0.69163E-01
msp                    -0.61507E-01     0.40672E-01
grade                   0.71592E-01     0.12613E-01
not_smsa               -0.12214         0.59017E-01
south                  -0.72293         0.58290E-01
scale                    1.3478         0.49969E-01
                                                             PROBABILITY
```

```
                                                               -----------
    endpoint 0            0.00000        FIXED          0.00000
    endpoint 1            0.21517E-01    0.54267E-02    0.21064E-01
```

The log likelihood of the homogeneous model of Task 1 is -7641.6559, and log likelihood of the random effects model of Task 2 is -7632.6474. The change in log likelihood over the Task 1 model is -2(-7641.6559+7632.6474)= 18.017. The sampling distribution of this test statistic is not chi-square with 2 df. Under the null hypothesis `endpoint 0 and 1` have the value 0, and they can only take values >0 under the alternative. The correct p value for this test statistics is obtained by dividing the naive p value of 18.017 for 2 degrees of freedom by 1/2, and so its clearly significant, suggesting that one or both are significant.

The estimate of `endpoint 0` is 0, suggesting that there is not a subgroup that will never be a union member. The estimate of the parameter for `endpoint 1` is small at 0.21517E-01 (S.E. 0.54267E-02), so that the probability of the upper endpoint is also small at 0.21064E-01 but it is significant and it does suggest that there is a subgroup of the population that will always be union members at this time.

The covariate parameter estimates of the model with endpoints are only slightly different to those of the model without, this is down to the fact that the magnitude of `endpoint 1` is small and that of `endpoint 0` is 0. The `scale` parameter of the model without endpoints is slightly larger because it is trying to include the stayers (extreme end of the distribution) as part of the Gaussian random effect distribution. It might be worth trying a nonparametric random effects distribution as an alternative to a continuous distribution with discrete endpoints.

22.2 Batch Script: `nlsunion_end.do`

```
log using nlsunion_end_s.log, replace
set more off
use nls
#delimit ;
sabre, data idcode year birth_yr age race msp nev_mar grade collgrad
            not_smsa c_city south union ttl_exp tenure ln_wage black age2
            ttl_exp2 tenure2;
sabre idcode year birth_yr age race msp nev_mar grade collgrad not_smsa
      c_city south union ttl_exp tenure ln_wage black age2 ttl_exp2 tenure2,
      read;
#delimit cr
sabre, case idcode
sabre, yvar union
sabre, link p
sabre, constant cons
sabre, quad a
sabre, mass 36
sabre, fit age age2 black msp grade not_smsa south cons
sabre, dis m
sabre, dis e
sabre, end b
sabre, fit age age2 black msp grade not_smsa south cons
sabre, dis m
sabre, dis e
log close
clear
```

```
exit
```

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

23.1 Relevant Results from `fish_s.log` and Discussion

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

Result/Discussion

```
Log likelihood =     -447.47621     on     246 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
cons                    -3.5349          0.64611
persons                 0.59934          0.14043
livebait                 1.4084          0.51517
scale                    1.9260          0.16693
```

Task 2. Re-estimate the same model but allow for lower endpoints. How much of an improvement in log likelihood do you get with the endpoints model? What happens to your inference on the covariates? How do you interpret the results of your preferred model?

Result/Discussion

```
Log likelihood =     -438.30927     on     245 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
cons                    -2.6703          0.56426
persons                 0.73530          0.11845
livebait                 1.5762          0.44179
scale                    1.1659          0.13378
                                                          PROBABILITY
                                                          ___________
endpoint 0              0.67121          0.14608          0.40163
```

The log likelihood of the homogeneous model of Task 1 is -447.47621, and log likelihood of the random effects model of Task 2 is -438.30927. The change in log likelihood over the Task 1 model is -2(-447.47621+438.30927)= 18.334. The sampling distribution of this test statistic is not chi-square with 2 df. Under the null hypothesis `endpoint 0` has the value 0, it can only take the value >0 under the alternative. The correct p value for this test statistics is obtained by dividing the naive p value of 18.334 for 1 degree of freedom by 1/2, and so its clearly significant, suggesting that there is a large subgroup who will never catch any fish, perhaps its because they do not fish.

The estimate of the parameter for the `endpoint 0` is large at 0.67121 (S.E. 0.14608), so that the probability of an endpoint is also large at 0.40163, it is very significant and it does suggest that there is a subgroup of the population that will never catch any fish.

The covariate parameter estimates are significant in both the Task 1 and Task 2 models. In the Task 2 model, the estimate of the `persons` effect has increased and its S.E has become smaller. The estimate of the `livebait` effect has also increased slightly and its S.E has also become smaller. Both models suggests that the use of `livebait` increases the rate at which fish are caught, and the larger the number of `persons` in the party the larger the rate at which fish are caught. The `scale` estimate is much larger in the model of Task 1 as it is trying to include the group that will never catch any fish (extreme left hand end of the latent distribution) as part of the Gaussian random effect distribution. It might be worth trying a nonparametric random effects distribution as an alternative to a continuous distribution with discrete endpoints.

23.2 Batch Script: `fish.do`

```
log using fish_s.log, replace
set more off
use fish
sabre, data nofish livebait camper persons child xb zg count id
sabre nofish livebait camper persons child xb zg count id, read
sabre, case id
sabre, yvar count
sabre, family p
sabre, constant cons
sabre, lfit persons livebait cons
sabre, dis m
sabre, dis e
sabre, quad a
sabre, mass 36
sabre, fit persons livebait cons
sabre, dis m
sabre, dis e
sabre, end l
sabre, fit persons livebait cons
sabre, dis m
sabre, dis e
log close
clear
exit
```

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

24.1 Relevant Results from `labour_s.log` and Discussion

Task 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 `mass 64`.

Result/Discussion

```
Log likelihood =     -3698.2985     on    7909 residual degrees of freedom

Parameter               Estimate         Std. Err.
___________________________________________________
cons                    -0.82912         0.13772
ft          (    1)       0.0000         ALIASED [I]
ft          (    2)      0.37129         0.11761
ft          (    3)      0.69983         0.11836
ft          (    4)      0.46031         0.11775
ft          (    5)      0.34388         0.11758
scale                     3.9658         0.15594
```

The covariate `ft(.)` is the factor for `t`, `ft(1)` is `ALIASED` as the model contains a constant.

Task 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 the results?

Result/Discussion

```
Log likelihood =     -3693.6887     on    7907 residual degrees of freedom

Parameter               Estimate         Std. Err.
___________________________________________________
cons                    -0.23907         0.18669
ft          (    1)       0.0000         ALIASED [I]
ft          (    2)      0.36716         0.11698
ft          (    3)      0.69568         0.11808
ft          (    4)      0.45568         0.11719
ft          (    5)      0.33996         0.11693
scale                     1.9485         0.39295
                                                           PROBABILITY
                                                           ___________
endpoint 0               0.41203         0.10310           0.24915
endpoint 1               0.24172         0.93774E-01       0.14616
```

The log likelihood of the homogeneous model of Task 1 is -3698.2985, and log likelihood of the random effects model of Task 2 is -3693.6887. The change in log likelihood over the Task 1 model is -2(-3698.2985+3693.6887)= 9.2196. The sampling distribution of this test statistic is not chi-square with 2 df. Under the null hypothesis `endpoint 0 and 1` have the value 0, and they can only take values >0 under the alternative. The correct p value for this test statistics is obtained by dividing the naive p value of 9.2196 for 2 degrees of freedom by 1/2, and so its clearly significant, suggesting that one or both are significant.

The estimate of the parameter for `endpoint 0` is 0.41203 (S.E 0.10310) suggesting that the probability that a randomly sampled woman from this population will never work over this time period is 0.24915. The estimate of the parameter for the `endpoint 1` is smaller at 0.24172 (S.E. 0.93774E-01), so that the probability of a randomly sampled female form this population will always work over this time period is 0.14616.

The parameter estimates of `ft(.)` are all significant, suggesting that the series is non stationary. The `scale` parameter of the model without endpoints is much larger because it is trying to include the both groups of stayers (both extreme ends of the latent distribution) as part of the Gaussian random effect distribution. It might be worth trying a nonparametric random effects distribution as an alternative to a continuous distribution with discrete endpoints.

24.2 Batch Script: `labour.do`

```
log using labour_s.log, replace
set more off
use labour
sabre, data case t y
sabre case t y, read
sabre, case case
sabre, yvar y
sabre, fac t ft
sabre, constant cons
sabre, lfit cons ft
sabre, dis m
sabre, dis e
sabre, quad a
sabre, mass 64
sabre, fit cons ft
sabre, dis m
sabre, dis e
sabre, end b
sabre, fit cons ft
sabre, dis m
sabre, dis e
log close
clear
exit
```

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

25.1 Conditional analysis: Relevant Results from `unionjmw_s.log` and Discussion

Task 1. Estimate a random effect probit model (adaptive quadrature, 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.

Result/Discussion

```
Log likelihood =     -1338.8321     on    3796 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
cons                    -1.3240         0.43605
union_1                  1.1275         0.10259
educ                   -0.19585E-01     0.35869E-01
black                   0.66836         0.18558
married                 0.17530         0.10904
marr81                  0.54328E-01     0.21341
marr82                  0.12027         0.25065
marr83                 -0.10103         0.25427
marr84                 -0.38317E-02     0.27284
marr85                  0.20568         0.25782
marr86                  0.13950         0.25941
marr87                 -0.30950         0.20259
d82                     0.51020E-02     0.11071
d83                    -0.11691         0.11477
d84                    -0.73547E-01     0.11643
d85                    -0.28268         0.11992
d86                    -0.31868         0.12205
d87                     0.67375E-01     0.11633
scale                    1.0919         0.10699
```

The parameter estimate for the lagged endogenous covariate `union_1` is the most significant effect in this conditional model. The estimates of the parameters for the time constant covariates `married` and `educ` are not significant, but `black` is. There is a lot of non stationarity effects in this model, but only the year dummy variables `d85` and `d86` are significant.

Task 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?

Result/Discussion

```
Log likelihood =     -1283.7471     on    3795 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
cons                    -1.6817         0.44298
union_1                  0.89739        0.92660E-01
union80                  1.4448         0.16437
educ                   -0.18453E-01     0.36230E-01
black                    0.52993        0.18371
married                  0.16892        0.11077
marr81                   0.42777E-01    0.21512
marr82                 -0.81286E-01     0.25313
marr83                 -0.88790E-01     0.25567
marr84                   0.26043E-01    0.27628
marr85                   0.39631        0.26087
marr86                   0.12489        0.26099
marr87                 -0.38636         0.20445
d82                      0.27602E-01    0.11368
d83                    -0.89635E-01     0.11753
d84                    -0.50365E-01     0.11913
d85                    -0.26696         0.12253
d86                    -0.31599         0.12449
d87                      0.73028E-01    0.11898
scale                    1.0765         0.90234E-01
```

The parameter estimate for `union_1` in Task 1 is 1.1275 (S.E. 0.10259). In task 2 this estimate is a lot smaller i.e. 0.89739 (S.E. 0.92660E-01). The estimate of the `scale` parameter hardly changes from Task1 to Task2. In Task 1 it is 1.0919 (S.E. 0.10699) and in Task 2 it is 1.0765 (S.E. 0.90234E-01). The estimates of the parameters for the time constant covariates have changed, `married` and `educ` are still not significant and the positive estimate on `black` is smaller. As in the Task 1 only the year dummy variables `d85` and `d86` are significant.

25.2 Joint analysis of the initial condition and subsequent responses: Relevant Results from `unionjmw_s.log` and Discussion

Task 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?

Result/Discussion

```
Log likelihood =     -1590.1430     on    4337 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
d1                    -0.58996         0.62227
d1_married             0.25759         0.20583
d1_educ               -0.48046E-01     0.52393E-01
d1_black               0.59148         0.26113
d2                     -1.2521         0.45364
d2_union_1             0.96357         0.87825E-01
d2_married             0.16569         0.10906
d2_educ               -0.27017E-01     0.37433E-01
d2_black               0.69899         0.19187
d2_marr81              0.97707E-01     0.19300
d2_marr82             -0.93949E-01     0.22448
d2_marr83             -0.89210E-01     0.22766
d2_marr84              0.36295E-01     0.24895
d2_marr85              0.38505         0.23111
d2_marr86              0.98316E-01     0.22917
d2_marr87             -0.35818         0.17973
d2_d82                 0.33469E-01     0.11200
d2_d83                -0.80935E-01     0.11563
d2_d84                -0.42037E-01     0.11717
d2_d85                -0.25302         0.12040
d2_d86                -0.29618         0.12218
d2_d87                 0.80604E-01     0.11719
scale                   1.1716         0.89832E-01
```

The parameter estimate for the lagged endogenous covariate `union_1` is 0.96357 (S.E. 0.87825E-01), it is the most significant covariate effect in this joint model. This estimate lies between those of the Task 1 and Task 2 conditional models. There is a very significant parameter estimate for the residual heterogeneity `scale,` which takes the value 1.1716 (S.E. 0.89832E-01) in this joint model. The only covariate effect that is significant in the model for the initial condition is `black`. The estimates of the parameters for the time constant covariates in the subsequent response model i.e. `married` and `educ` are still not significant and the positive estimate on `black` is larger than previously. As in the Task 1 and Task 2 conditional models, non of the `marr81-marr86` effects are significant, but `marr87` now is now marginally significant. As before, the year dummy variables `d85` and `d86` are significant.

Task 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?

Result/Discussion

```
Log likelihood =     -1587.3937     on    4336 residual degrees of freedom

Parameter              Estimate        Std. Err.
___________________________________________________
d1                    -0.55996         0.55785
d1_married             0.23441         0.18924
d1_educ               -0.40286E-01     0.46985E-01
d1_black               0.52854         0.23547
d2                     -1.2616         0.49391
d2_union_1             0.89734         0.92530E-01
d2_married             0.16901         0.11093
d2_educ               -0.30145E-01     0.40841E-01
d2_black               0.74873         0.21034
d2_marr81              0.10080         0.21806
d2_marr82             -0.79352E-01     0.25414
d2_marr83             -0.91932E-01     0.25750
d2_marr84              0.31681E-01     0.28034
d2_marr85              0.39320         0.26147
d2_marr86              0.11828         0.26002
d2_marr87             -0.38018         0.20383
d2_d82                 0.29233E-01     0.11386
d2_d83                -0.87934E-01     0.11768
d2_d84                -0.48132E-01     0.11928
d2_d85                -0.26486         0.12262
d2_d86                -0.31378         0.12458
d2_d87                 0.75523E-01     0.11921
scale1                 0.93682         0.11943
scale2                  1.2928         0.10895
```

The log likelihood of the common random effect model of Task 3 is -1590.1430 and log likelihood of the random effects model of Task 4 is -1587.3937. The change in log likelihood over the Task 3 model is -2(-1590.1430+1587.3937)= 5.4986. The sampling distribution of this test statistic is chi-square with 1 df. Under the null hypothesis `scale1` and 2 are equal, The test statistic is clearly significant, suggesting that `scale1` and `scale2` are significantly different from each other.

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

Result/Discussion

```
Log likelihood =     -1587.3902     on    4335 residual degrees of freedom

Parameter              Estimate        Std. Err.
___________________________________________________
d1                    -0.55091         0.54565
```

```
d1_married              0.22934          0.19315
d1_educ                -0.37589E-01      0.54714E-01
d1_black                0.50766          0.32984
d2                     -1.2900           0.59611
d2_union_1              0.89724          0.92550E-01
d2_union80              0.99365E-01       1.2048
d2_married              0.16896          0.11090
d2_educ                -0.29059E-01      0.42171E-01
d2_black                0.73161          0.29121
d2_marr81               0.96561E-01      0.22362
d2_marr82              -0.80274E-01      0.25416
d2_marr83              -0.90811E-01      0.25754
d2_marr84               0.30503E-01      0.27999
d2_marr85               0.39368          0.26141
d2_marr86               0.11981          0.26070
d2_marr87              -0.38154          0.20444
d2_d82                  0.29096E-01      0.11385
d2_d83                 -0.88060E-01      0.11766
d2_d84                 -0.48324E-01      0.11928
d2_d85                 -0.26502          0.12262
d2_d86                 -0.31395          0.12458
d2_d87                  0.75265E-01      0.11921
scale1                  0.85413          0.98069
scale2                   1.2631          0.36167
```

The log likelihood of the common random effect but different scales model of Task 4 is -1587.3937 and log likelihood of the model of Task 5 is -1587.3902. The change in log likelihood over the Task 4 model is -2(-1587.3937+1587.3902)= 0.007. The sampling distribution of this test statistic is chi-square with 1 df. Under the null hypothesis `d2_union80=0`. The test statistic is clearly not significant. The same result is given by the z statistic for the parameter estimate of `d2_union80` which is 0.099365/1.2048= 8.2474×10^{-2}.

25.3 Batch Script: `unionjmw.do`

```
log using unionjmw_s.log, replace
set more off
use unionjmw1
#delimit ;
sabre, data nr year black married educ union d81 d82 d83 d84 d85 d86 d87
            union80 union_1 marravg educu80 marr81 marr82 marr83 marr84
            marr85 marr86 marr87;
sabre nr year black married educ union d81 d82 d83 d84 d85 d86 d87 union80
      union_1 marravg educu80 marr81 marr82 marr83 marr84 marr85 marr86
      marr87, read;
#delimit cr
sabre, case nr
sabre, yvar union
sabre, link p
sabre, constant cons
sabre, quad a
sabre, mass 24
#delimit ;
```

```
sabre, fit union_1 educ black married marr81 marr82 marr83 marr84 marr85
           marr86 marr87 d82 d83 d84 d85 d86 d87 cons;
#delimit cr
sabre, dis m
sabre, dis e
#delimit ;
sabre, fit union_1 union80 educ black married marr81 marr82 marr83 marr84
           marr85 marr86 marr87 d82 d83 d84 d85 d86 d87 cons;
#delimit cr
sabre, dis m
sabre, dis e
clear
use unionjmw2
#delimit ;
sabre, data nr year black married educ union d81 d82 d83 d84 d85 d86 d87
            union80 union_1 marravg educu80 marr81 marr82 marr83 marr84
            marr85 marr86 marr87 d d1 d2;
sabre nr year black married educ union d81 d82 d83 d84 d85 d86 d87 union80
      union_1 marravg educu80 marr81 marr82 marr83 marr84 marr85 marr86
      marr87 d d1 d2, read;
#delimit cr
sabre, case nr
sabre, yvar union
sabre, rvar d
sabre, link p
sabre, trans d1_educ d1 * educ
sabre, trans d1_black d1 * black
sabre, trans d1_married d1 * married
sabre, trans d2_union_1 d2 * union_1
sabre, trans d2_union80 d2 * union80
sabre, trans d2_educ d2 * educ
sabre, trans d2_black d2 * black
sabre, trans d2_married d2 * married
sabre, trans d2_marr81 d2 * marr81
sabre, trans d2_marr82 d2 * marr82
sabre, trans d2_marr83 d2 * marr83
sabre, trans d2_marr84 d2 * marr84
sabre, trans d2_marr85 d2 * marr85
sabre, trans d2_marr86 d2 * marr86
sabre, trans d2_marr87 d2 * marr87
sabre, trans d2_d82 d2 * d82
sabre, trans d2_d83 d2 * d83
sabre, trans d2_d84 d2 * d84
sabre, trans d2_d85 d2 * d85
sabre, trans d2_d86 d2 * d86
sabre, trans d2_d87 d2 * d87
sabre, quad a
sabre, mass 24
#delimit ;
sabre, fit d1 d1_married d1_educ d1_black
           d2 d2_union_1 d2_married d2_educ d2_black d2_marr81 d2_marr82
           d2_marr83 d2_marr84 d2_marr85 d2_marr86 d2_marr87 d2_d82 d2_d83
           d2_d84 d2_d85 d2_d86 d2_d87;
#delimit cr
sabre, dis m
sabre, dis e
sabre, depend y
sabre, nvar 4
#delimit ;
sabre, fit d1 d1_married d1_educ d1_black
           d2 d2_union_1 d2_married d2_educ d2_black d2_marr81 d2_marr82
           d2_marr83 d2_marr84 d2_marr85 d2_marr86 d2_marr87 d2_d82 d2_d83
```

```
          d2_d84 d2_d85 d2_d86 d2_d87;
#delimit cr
sabre, dis m
sabre, dis e
sabre, nvar 4
#delimit ;
sabre, fit d1 d1_married d1_educ d1_black
          d2 d2_union_1 d2_union80 d2_married d2_educ d2_black d2_marr81
          d2_marr82 d2_marr83 d2_marr84 d2_marr85 d2_marr86 d2_marr87
          d2_d82 d2_d83 d2_d84 d2_d85 d2_d86 d2_d87;
#delimit cr
sabre, dis m
sabre, dis e
log close
clear
exit
```

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

26.1 Conditional analysis: Relevant Results from `unionred_s.log` and Discussion

Task 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.

Result/Discussion

```
Log likelihood =     -1561.1661     on    3989 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
cons                  -0.12753          0.39251
lagunion                1.1723          0.14108
age                   -0.15189E-01      0.84733E-02
grade                 -0.38049E-01      0.20260E-01
southxt               -0.27348E-01      0.67395E-02
scale                   1.0210          0.15065
```

The parameter estimate for the lagged endogenous covariate (`lagunion)` is the most significant effect in this conditional random effects model, its z statistic is 1.1723/0.14108= 8.309 5. The estimates of the parameters for `grade` and `age` are marginally significant, but the estimates of `southxt` is very significant.

Task 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.

Result/Discussion

```
Log likelihood =     -1440.9676     on    3988 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
cons                  -0.55370E-01      0.41636
lagunion               0.61315          0.97749E-01
baseunion               2.0856          0.18478
age                   -0.23876E-01      0.91305E-02
grade                 -0.58040E-01      0.22610E-01
southxt               -0.15529E-01      0.71251E-02
scale                   1.1519          0.94868E-01
```

The parameter estimate for `lagunion` in Task 1 is 1.1723 (S.E. 0.14108). In task 2 this estimate is a lot smaller i.e. 0.61315 (S.E. 0.97749E-01). The estimate of the `scale` parameter hardly changes from Task 1 to Task 2. In Task 1 it is 1.0210 (S.E. 0.15065) and in Task 2 it is 1.1519 (S.E. 0.94868E-01). The estimates for the other covariate parameters have changed. The estimates of the parameters for `grade` and `age` are now significant, but the estimates of `southxt` is now of marginal significance, suggesting that the very significant endogenous covariate `baseunion` is correlated with these explanatory covariates.

26.2 Joint analysis of the initial condition and subsequent responses: Relevant Results from `unionred_s.log` and Discussion

Task 3. Estimate a common random effect common scale joint probit model (adaptive 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?

Result/Discussion

```
Log likelihood =     -1859.3298     on    4783 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
d1                      -1.2135         0.87794
d1_age                   0.15555E-01    0.24851E-01
d1_grade               -0.63505E-02     0.35847E-01
d1_southxt             -0.96174E-01     0.20873E-01
d1_not_smsa            -0.44161         0.16998
d2                       0.69683E-01    0.44656
d2_lagunion              0.68544        0.90929E-01
d2_age                 -0.15415E-01     0.92712E-02
d2_grade               -0.49664E-01     0.25326E-01
d2_southxt             -0.33817E-01     0.76453E-02
scale                    1.4361         0.10073
```

The parameter estimate for the lagged endogenous covariate (`d2_lagunion)` is 0.68544 (S.E. 0.90929E-01), it is the most significant covariate effect in this joint model. This estimate lies between those of the Task 1 and Task 2 conditional models. There is a very significant parameter estimate for the residual heterogeneity `scale,` which takes the value 1.4361 (S.E. 0.10073). This estimate of the `scale` effect is larger than the estimates of Task 1 and Task 2. The only covariate effects that are significant in the model for the initial condition are: `d1_southxt` and `d1_not_smsa`. The estimates of the parameters for the time constant covariates in the subsequent response model, i.e. `d2_grade,` `d2_southxt` are significant. The estimate `d2_age` is not significant.

Task 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?

Result/Discussion

```
Log likelihood =     -1858.7970     on    4782 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
d1                      -1.2135         0.83951
d1_age                   0.16495E-01    0.23826E-01
d1_grade                -0.51061E-02    0.34065E-01
d1_southxt              -0.91276E-01    0.20370E-01
d1_not_smsa             -0.41669        0.16479
d2                       0.11430        0.46088
d2_lagunion              0.64705        0.98257E-01
d2_age                  -0.16227E-01    0.94731E-02
d2_grade                -0.52032E-01    0.26467E-01
d2_southxt              -0.34468E-01    0.79324E-02
scale1                   1.3189         0.14238
scale2                   1.5062         0.12400
```

The log likelihood of the common random effect model of Task 3 is -1859.3298 and log likelihood of the random effects model of Task 4 is -1858.7970. The change in log likelihood over the Task 3 model is -2(-1859.3298+1858.7970)= 1.065 6. The sampling distribution of this test statistic is chi-square with 1 df. Under the null hypothesis `scale1` and 2 are equal, The test statistic is clearly not significant, suggesting that `scale1` and `scale2` are not significantly different from each other.

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

Result/Discussion

```
Log likelihood =     -1858.7970     on    4782 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
d1                      -1.3255         0.92173
d1_age                   0.18018E-01    0.25485E-01
d1_grade                -0.55778E-02    0.37603E-01
d1_southxt              -0.99705E-01    0.23454E-01
d1_not_smsa             -0.45517        0.19332
```

```
d2                       0.11430          0.45735
d2_lagunion              0.64705          0.99895E-01
d2_age                  -0.16227E-01      0.96947E-02
d2_grade                -0.52032E-01      0.25438E-01
d2_southxt              -0.34468E-01      0.82239E-02
scale                    1.5062           0.12352
corr                     0.95647          0.40383E-01
```

There is not much difference between the log likelihood and results and those of Task 3 (log likelihood -1859.3298) or Task 4 (log likelihood -1858.7970). This is reinforced by the fact that the 95% confidence interval on `corr` includes 1, a value which gives the common random effect model of Task 3 and the estimated different scales model of Task 4.

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

Result/Discussion

```
Log likelihood =     -1849.0718     on    4781 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
d1                      -2.6975           0.94777
d1_age                   0.68087E-01      0.26635E-01
d1_grade                 0.32122E-02      0.34305E-01
d1_southxt              -0.10413          0.22391E-01
d1_not_smsa             -0.43624          0.17741
d2                      -0.81790E-01      0.44251
d2_lagunion              0.61259          0.10019
d2_baseunion             2.5607           0.79879
d2_age                  -0.26334E-01      0.98441E-02
d2_grade                -0.59834E-01      0.22166E-01
d2_southxt              -0.11618E-01      0.94130E-02
scale                    1.1707           0.10772
corr                    -0.31741          0.51614
```

The log likelihood of the common scale different random effect model of Task 5 is -1858.7970 and log likelihood of the model of Task 6 is -1849.0718. The change in log likelihood over the Task 5 model is -2(-1858.7970+1849.0718)= 19.45. The sampling distribution of this test statistic is chi-square with 1 df. Under the null hypothesis `d2_baseunion=0`. The test statistic for `d2_baseunion` not equal to 0 is clearly significant. The same result is given by the z statistic for the parameter estimate of `d2_baseunion` which is 2.5607/0.79879= 3.2057.

In this bivariate model `corr` is estimated to be negative but non significant, implying independence between the initial condition and the subsequent responses, perhaps the Task2 model is a reasonable representation of the data.

26.3 Batch Script: `unionred.do`

```
log using unionred_s.log, replace
set more off
use unionred1
#delimit ;
sabre, data idcode year age grade not_smsa south union t0 southXt black tper
            lagunion d d1 d2 baseunion;
sabre idcode year age grade not_smsa south union t0 southXt black tper
             lagunion d d1 d2 baseunion, read;
#delimit cr
sabre, case idcode
sabre, yvar union
sabre, link p
sabre, constant cons
sabre, quad a
sabre, mass 16
sabre, fit lagunion age grade southXt cons
sabre, dis m
sabre, dis e
sabre, fit lagunion baseunion age grade southXt cons
sabre, dis m
sabre, dis e
clear
use unionred2
#delimit ;
sabre, data idcode year age grade not_smsa south union t0 southXt black tper
            lagunion d d1 d2 baseunion;
sabre idcode year age grade not_smsa south union t0 southXt black tper
      lagunion d d1 d2 baseunion, read;
#delimit cr
sabre, case idcode
sabre, yvar union
sabre, rvar d
sabre, link p
sabre, trans d1_age d1 * age
sabre, trans d1_grade d1 * grade
sabre, trans d1_southXt d1 * southXt
sabre, trans d1_not_smsa d1 * not_smsa
sabre, trans d2_lagunion d2 * lagunion
sabre, trans d2_baseunion d2 * baseunion
sabre, trans d2_age d2 * age
sabre, trans d2_grade d2 * grade
sabre, trans d2_southXt d2 * southXt
sabre, quad a
sabre, mass 24
#delimit ;
sabre, fit d1 d1_age d1_grade d1_southXt d1_not_smsa
           d2 d2_lagunion d2_age d2_grade d2_southXt;
#delimit cr
sabre, dis m
sabre, dis e
sabre, depend y
sabre, mass 32
sabre, nvar 5
#delimit ;
sabre, fit d1 d1_age d1_grade d1_southXt d1_not_smsa
           d2 d2_lagunion d2_age d2_grade d2_southXt;
#delimit cr
sabre, dis m
sabre, dis e
sabre, model b
sabre, eqscale y
```

```
sabre, der1 y
sabre, mass first=24 second=24
sabre, nvar 5
#delimit ;
sabre, fit d1 d1_age d1_grade d1_southXt d1_not_smsa
           d2 d2_lagunion d2_age d2_grade d2_southXt;
#delimit cr
sabre, dis m
sabre, dis e
sabre, nvar 5
#delimit ;
sabre, fit d1 d1_age d1_grade d1_southXt d1_not_smsa
           d2 d2_lagunion d2_baseunion d2_age d2_grade d2_southXt;
#delimit cr
sabre, dis m
sabre, dis e
log close
clear
exit
```

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

27.1 Conditional analysis: Relevant Results from `wemp_base_s.log` and Discussion

Task 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..

Result/Discussion

```
Log likelihood =     -384.71153     on    1268 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
cons                  -0.84840          0.25399
ylag                    3.7180          0.25145
mune                   -1.6654          0.44273
und5                   -1.0786          0.28686
age                    0.79040E-03      0.16505E-01
scale                  0.87551          0.25075
```

The parameter estimate for the lagged endogenous covariate (`ylag`) is the most significant effect in this conditional random effects model, its z statistic is 3.7180/0.25145= 14.786. The estimates of the parameters for `mune` and `und5` are very significant, but the estimate of `age` is not significant.

Task 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?

Result/Discussion

```
Log likelihood =     -380.63889     on    1267 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
cons                   -1.1012          0.26233
ylag                    3.3566          0.26986
ybase                  0.91324          0.35759
mune                   -1.7769          0.46858
und5                   -1.1307          0.29507
age                    0.34266E-03      0.17862E-01
scale                   1.0665          0.24790
```

The parameter estimate for `ylag` in Task 1 is 3.7180 (S.E. 0.25145). In task 2 this estimate is smaller i.e. 3.3566 (S.E. 0.26986). The estimate of the `scale` parameter is larger in the Task 2 model than it is in the Task 2 model. In Task 1 it is 0.87551 (S.E. 0.25075) and in Task 2 it is 1.0665 (S.E. 0.24790). The estimates for the other covariate parameters have changed slightly, but the pattern of significance is the same, suggesting that the significant endogenous covariate `ybase` is only lightly correlated with these explanatory covariates.

27.2 Joint analysis of the initial condition and subsequent responses: Relevant Results from `wemp_base_s.log` and Discussion

Task 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?

Result/Discussion

```
Log likelihood =     -463.56628     on    1415 residual degrees of freedom

Parameter              Estimate        Std. Err.
___________________________________________________
r1                       1.5314         0.32754
r1_mune                 -1.5048         0.96871
r1_und5                 -2.4403         0.49140
r1_age                  0.39628E-02     0.29897E-01
r2                     -0.57860         0.26726
r2_ylag                  3.3681         0.26379
r2_mune                 -1.9178         0.47149
r2_und5                 -1.1457         0.29263
r2_age                  0.52903E-02     0.18133E-01
scale                    1.1572         0.23703
```

The parameter estimate for the lagged endogenous covariate (`r2_ylag`) is 3.3681 (S.E. 0.26379), it is the most significant covariate effect in this joint model. This estimate lies between those of the Task 1 and Task 2 conditional models. There is a very significant parameter estimate for the residual heterogeneity `scale,` which takes the value 1.1572 (S.E. 0.23703). This estimate of the `scale` effect is larger than the estimates of Task 1 and Task 2. The only covariate effect that is significant in the model for the initial condition is `r1_und5`. The estimates of the parameters for the time constant covariates in the subsequent response model, i.e. `r2_mune, r2_und5` are significant. The estimate `r2_age` is not significant.

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

Result/Discussion

```
Log likelihood =     -463.55824     on    1414 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
r1                       1.5554         0.38557
r1_mune                 -1.5126         0.98150
r1_und5                 -2.4821         0.60308
r1_age                  0.40872E-02     0.30345E-01
r2                     -0.58392         0.26949
r2_ylag                  3.3699         0.26373
r2_mune                 -1.9135         0.47098
r2_und5                 -1.1415         0.29341
r2_age                  0.51288E-02     0.18080E-01
scale1                   1.2085         0.47715
scale2                   1.1424         0.26382
```

The estimates of `scale1` and `scale2` look very similar. The log likelihood of the common random effect model of Task 3 is -463.56628 and log likelihood of the random effects model of Task 4 is -463.55824. The change in log likelihood over the Task 3 model is -2(-463.56628+463.55824)= 0.016 08. The sampling distribution of this test statistic is chi-square with 1 df. Under the null hypothesis `scale1` and 2 are equal. The test statistic is clearly not significant, suggesting that `scale1` and `scale2` are not significantly different from each other.

Task 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?

Result/Discussion

```
Log likelihood =     -547.21951     on    1414 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
r1                       1.3616         0.32490
r1_mune                 -1.3711         0.90300
r1_und5                 -2.1719         0.50052
r1_age                  0.38719E-02     0.26628E-01
r2                      0.77068         0.70389
```

```
r2_ybase               2.1017          1.0935
r2_mune               -2.5860          0.54281
r2_und5               -2.0783          0.30095
r2_age                0.21867E-01      0.24741E-01
scale1                0.70483          0.55268
scale2                 2.7334          0.34122
```

The estimates of `scale1` and `scale2` now seem to be very different, in fact `scale1` looks to be non significant, perhaps the inclusion of `r2_ybase` in the model for the subsequent responses has captured the dependence between the two sub models. The log likelihood of the Task 5 model is -547.21951 which is much poorer than the model of Task 4 is -463.55824. The Task 4 and 5 models are not nested, so we can not formally compare the two models using a likelihood ratio test.

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

Result/Discussion

```
Log likelihood =     -463.52580     on    1413 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
r1                     1.4748          0.45687
r1_mune               -1.4938          0.94264
r1_und5               -2.3532          0.72203
r1_age                0.30407E-02      0.28941E-01
r2                   -0.66531          0.41447
r2_ylag                3.3563          0.26850
r2_ybase              0.15846          0.62399
r2_mune               -1.8982          0.47553
r2_und5               -1.1411          0.29391
r2_age                0.43770E-02      0.18278E-01
scale1                 1.0035          0.87665
scale2                 1.1246          0.26678
```

The log likelihood of the common scale different random effect model of Task 5 is -547.21951 and log likelihood of the model of Task 6 is -463.52580. The change in log likelihood over the Task 5 model is -2(-547.21951+463.52580)= 167.39 The sampling distribution of this test statistic is chi-square with 1 df. Under the null hypothesis `r2_ylag=0`. The test statistic for `r2_ylag` not equal to 0 is clearly significant. The same result is given by the z statistic for the parameter estimate of `r2_ylag` which is 3.3563/0.26850= 12.5. The z statistic for the parameter estimate of `r2_ybase` is 0.15846/0.62399= 0.253 95 which is not significant. The estimates of `scale1` and `scale2` look very similar, as in the Task 4 model.

Task 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).

Result/Discussion

```
Log likelihood =     -463.53052     on    1413 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
r1                       1.5262          0.35220
r1_mune                 -1.5241          0.86212
r1_und5                 -2.4372          0.49228
r1_age                  0.33644E-02      0.33696E-01
r2                     -0.65442          0.40181
r2_ylag                  3.3582          0.27252
r2_ybase                0.13621          0.64180
r2_mune                 -1.8994          0.40353
r2_und5                 -1.1409          0.26281
r2_age                  0.44463E-02      0.17992E-01
scale                    1.1244          0.24811
corr                    0.94690          0.60503
```

There is not much difference between the log likelihood and results of the Task 4 model (log likelihood -463.55824), the Task 6 model (log likelihood -463.52580) and those of the Task 7 model (log likelihood -463.53052). This is reinforced by the fact that the estimate of `r2_ybase` is not significant in the Task 7 model and the 95% confidence interval on `corr` includes 1, a value which gives the common random effect model of Task 4 and the estimated different scales model of Task 6.

Task 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?

Result/Discussion

The results obtained for the various models (Task 4, 5, 6, 7) on the covariates and role of employment status in the previous year are very similar. In the joint models of Tasks 6 and 7 which contain both `r2_ylag` and `r2_ybase`, `r2_ybase` is not significant. The estimate of the state dependence effect (`r2_ylag`) in the Task 7 model is 3.3582 (S.E. 0.27252), it has a z statistic of 3.3582/0.27252= 12. 323, which is very significant. Similar inference occurs in the Task 4, and 6 models. The 95% confidence interval on the `scale` parameter estimate does not include 0, suggesting the presence of residual heterogeneity. Both state dependence and unobserved heterogeneity present in this data.

27.3 Batch Script: `wemp_base.do`

```
log using wemp_base_s.log, replace
set more off
use wemp_base1
sabre, data case femp mune time und1 und5 age d d1 d0 ylag ybase r r1 r2
sabre case femp mune time und1 und5 age d d1 d0 ylag ybase r r1 r2, read
sabre, case case
sabre, yvar femp
sabre, constant cons
sabre, quad a
sabre, mass 12
sabre, fit ylag mune und5 age cons
sabre, dis m
sabre, dis e
sabre, fit ylag ybase mune und5 age cons
sabre, dis m
sabre, dis e
clear
use wemp_base2
sabre, data case femp mune time und1 und5 age d d1 d0 ylag ybase r r1 r2
sabre case femp mune time und1 und5 age d d1 d0 ylag ybase r r1 r2, read
sabre, case case
sabre, yvar femp
sabre, rvar r
sabre, trans r1_mune r1 * mune
sabre, trans r1_und5 r1 * und5
sabre, trans r1_age r1 * age
sabre, trans r2_ylag r2 * ylag
sabre, trans r2_ybase r2 * ybase
sabre, trans r2_mune r2 * mune
sabre, trans r2_und5 r2 * und5
sabre, trans r2_age r2 * age
sabre, quad a
sabre, mass 12
sabre, fit r1 r1_mune r1_und5 r1_age r2 r2_ylag r2_mune r2_und5 r2_age
sabre, dis m
sabre, dis e
sabre, depend y
sabre, nvar 4
sabre, fit r1 r1_mune r1_und5 r1_age r2 r2_ylag r2_mune r2_und5 r2_age
sabre, dis m
sabre, dis e
sabre, nvar 4
sabre, fit r1 r1_mune r1_und5 r1_age r2 r2_ybase r2_mune r2_und5 r2_age
sabre, dis m
sabre, dis e
sabre, nvar 4
#delimit ;
sabre, fit r1 r1_mune r1_und5 r1_age
           r2 r2_ylag r2_ybase r2_mune r2_und5 r2_age;
#delimit cr
sabre, dis m
sabre, dis e
sabre, depend n
sabre, model b
sabre, eqscale y
sabre, der1 y
sabre, mass first=24 second=24
sabre, nvar 4
#delimit ;
sabre, fit r1 r1_mune r1_und5 r1_age
           r2 r2_ylag r2_ybase r2_mune r2_und5 r2_age;
```

```
#delimit cr
sabre, dis m
sabre, dis e
log close
clear
exit
```

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

28.1 Relevant Results from `patents_s.log` and Discussion

Task 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, as follows:

- `trans r1_logr r1 * logr`
- `trans r1_logk r1 * logk`
- `trans r1_scisect r1 * scisect`
- `trans r2_logr r2 * logr`
- `trans r2_logk r2 * logk`
- `trans r2_scisect r2 * scisect`
- `trans r2_year3 r2 * year3`
- `trans r2_year4 r2 * year4`
- `trans r2_year5 r2 * year5`
- `trans r2_pat1 r2 * pat1`
- `trans r2_base r2 * base`

Task 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 with `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?

Result/Discussion

(a) Common random effect model to baseline and subsequent responses without endogenous covariates.

```
Log likelihood =     -5109.3189     on    1668 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
r1                     -0.31596         0.17375
```

```
r1_logr                 0.52562          0.37290E-01
r1_logk                 0.33700          0.41101E-01
r1_scisect              0.50912          0.12782
r2                     -0.43888          0.16764
r2_logr                 0.48243          0.34783E-01
r2_logk                 0.37341          0.39376E-01
r2_scisect              0.53284          0.12622
r2_year3               -0.76923E-02      0.12885E-01
r2_year4               -0.13744          0.13595E-01
r2_year5               -0.18812          0.14428E-01
scale                    1.0262          0.49693E-01
```

(b) Common random effect model to baseline and subsequent responses with `pat1`.

```
Log likelihood =     -5103.4358     on    1667 residual degrees of freedom

Parameter               Estimate         Std. Err.
---------------------------------------------------
r1                     -0.31642          0.17251
r1_logr                 0.54497          0.37556E-01
r1_logk                 0.33078          0.40870E-01
r1_scisect              0.49212          0.12686
r2                     -0.39311          0.16681
r2_pat1                 0.30541E-03      0.89147E-04
r2_logr                 0.48773          0.34637E-01
r2_logk                 0.35968          0.39291E-01
r2_scisect              0.51490          0.12524
r2_year3               -0.62285E-02      0.12892E-01
r2_year4               -0.13618          0.13596E-01
r2_year5               -0.18114          0.14561E-01
scale                    1.0166          0.49293E-01
```

(c) Common random effect model to baseline and subsequent responses with `pat1` and `base`.

```
Log likelihood =     -5010.8108     on    1666 residual degrees of freedom

Parameter               Estimate         Std. Err.
---------------------------------------------------
r1                     -0.32317          0.17456
r1_logr                 0.49949          0.37931E-01
r1_logk                 0.34851          0.41333E-01
r1_scisect              0.52671          0.12868
r2                     -0.48806          0.16923
r2_pat1                 0.20237E-02      0.15990E-03
r2_base                -0.22065E-02      0.16540E-03
r2_logr                 0.48395          0.34898E-01
r2_logk                 0.38086          0.39783E-01
```

```
r2_scisect             0.54412          0.12710
r2_year3               0.37597E-02      0.12923E-01
r2_year4              -0.12636          0.13631E-01
r2_year5              -0.14101          0.14850E-01
scale                   1.0331          0.50124E-01
```

The log likelihood improves at each step, (a) -5109.3189, (b) -5103.4358, (c) -5010.8108. Each improvement has a significant chi square statistic (not shown), suggesting that both the endogenous covariates `pat1` and `base` are significant. The biggest improvement is between models b and c.

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

Result/Discussion

(a) 1 factor random effect model to baseline and subsequent responses without endogenous covariates.

```
Log likelihood =     -5108.0097     on    1667 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
r1                    -0.26999          0.17237
r1_logr                0.52802          0.36580E-01
r1_logk                0.33165          0.40372E-01
r1_scisect             0.49814          0.12521
r2                    -0.43901          0.16837
r2_logr                0.48698          0.35072E-01
r2_logk                0.37082          0.39589E-01
r2_scisect             0.52956          0.12679
r2_year3              -0.78688E-02      0.12886E-01
r2_year4              -0.13788          0.13602E-01
r2_year5              -0.18886          0.14447E-01
scale1                  1.0032          0.50564E-01
scale2                  1.0306          0.49942E-01
```

(b) 1 factor random effect model to baseline and subsequent responses with `pat1`.

```
Log likelihood =     -5103.4351     on    1666 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
r1                    -0.31766          0.17572
r1_logr                0.54503          0.37613E-01
r1_logk                0.33088          0.40981E-01
```

```
r1_scisect             0.49230         0.12702
r2                    -0.39283         0.16695
r2_pat1                0.30716E-03     0.10051E-03
r2_logr                0.48764         0.34711E-01
r2_logk                0.35967         0.39286E-01
r2_scisect             0.51488         0.12522
r2_year3              -0.62152E-02     0.12896E-01
r2_year4              -0.13616         0.13605E-01
r2_year5              -0.18108         0.14645E-01
scale1                 1.0172          0.51527E-01
scale2                 1.0164          0.49497E-01
```

(c) 1 factor random effect model to baseline and subsequent responses with `pat1` and `base`.

```
Log likelihood =     -5004.1494     on    1665 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
r1                    -0.19313         0.16899
r1_logr                0.48936         0.36085E-01
r1_logk                0.33873         0.39252E-01
r1_scisect             0.51002         0.12193
r2                    -0.53394         0.17326
r2_pat1                0.19393E-02     0.16226E-03
r2_base               -0.24280E-02     0.18114E-03
r2_logr                0.49286         0.35658E-01
r2_logk                0.38638         0.40635E-01
r2_scisect             0.55199         0.12987
r2_year3               0.27664E-02     0.12928E-01
r2_year4              -0.12795         0.13653E-01
r2_year5              -0.14462         0.14929E-01
scale1                 0.97767         0.49653E-01
scale2                 1.0560          0.51557E-01
```

The log likelihood improves at each step, (a) -5108.0097, (b) -5103.4351, (c) -5004.1494. Each improvement has a significant chi square statistic (not shown), suggesting that both the endogenous covariates `pat1` and `base` are significant. The biggest improvement is between models b and c.

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

Result/Discussion

(a) Bivariate random effect model to baseline and subsequent responses without endogenous covariates.

```
Log likelihood =     -4994.0714     on    1666 residual degrees of freedom

Parameter              Estimate         Std. Err.
---------------------------------------------------
r1                    -0.17586          0.17752
r1_logr                0.56408          0.42068E-01
r1_logk                0.30412          0.43150E-01
r1_scisect             0.45684          0.12411
r2                    -0.34140          0.17148
r2_logr                0.53246          0.37611E-01
r2_logk                0.33564          0.40939E-01
r2_scisect             0.47559          0.12796
r2_year3              -0.94811E-02      0.12894E-01
r2_year4              -0.14219          0.13657E-01
r2_year5              -0.19627          0.14609E-01
scale1                 0.95748          0.50841E-01
scale2                  1.0307          0.49924E-01
corr                   0.97055          0.65365E-02
```

(b) Bivariate random effect model to baseline and subsequent responses with `pat1`.

```
Log likelihood =     -4964.8702     on    1665 residual degrees of freedom

Parameter              Estimate         Std. Err.
---------------------------------------------------
r1                    -0.21356          0.18249
r1_logr                0.59108          0.43822E-01
r1_logk                0.29866          0.44499E-01
r1_scisect             0.44247          0.12650
r2                    -0.24339          0.16758
r2_pat1                0.11669E-02      0.15559E-03
r2_logr                0.52925          0.37078E-01
r2_logk                0.30651          0.40216E-01
r2_scisect             0.43689          0.12443
r2_year3              -0.34130E-02      0.12921E-01
r2_year4              -0.13639          0.13671E-01
r2_year5              -0.16896          0.15006E-01
scale1                 0.96908          0.52022E-01
scale2                 0.99743          0.48700E-01
corr                   0.96375          0.76988E-02
```

(c) Bivariate random effect model to baseline and subsequent responses with `pat1` and `base`.

```
Log likelihood =     -4954.9182     on    1664 residual degrees of freedom

Parameter              Estimate         Std. Err.
----------------------------------------------------
r1                    -0.17134          0.17566
r1_logr                0.55253          0.41689E-01
r1_logk                0.30843          0.42315E-01
r1_scisect             0.46404          0.12315
r2                    -0.37801          0.17492
r2_pat1                0.14635E-02      0.16918E-03
r2_base               -0.16876E-02      0.34728E-03
r2_logr                0.53407          0.37547E-01
r2_logk                0.34224          0.41798E-01
r2_scisect             0.48522          0.12904
r2_year3              -0.19797E-02      0.12931E-01
r2_year4              -0.13535          0.13689E-01
r2_year5              -0.16317          0.15090E-01
scale1                 0.95599          0.50521E-01
scale2                  1.0372          0.51278E-01
corr                   0.97859          0.55435E-02
```

The log likelihood improves at each step, (a) -4994.0714, (b) -4964.8702, (c) -4954.9182. Each improvement has a significant chi square statistic (not shown), suggesting that both the endogenous covariates `pat1` and `base` are significant. The biggest improvement is between models a and b.

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

Result/Discussion

In all 3 Tasks the preferred model is model c. All 3 Tasks suggest the presence of a positive effect for the lagged response for the number of patents applied for during the previous year. We are unaware of anyone else who has found this effect in this data. The three models of Task 2 and 3 are very similar. The Task 4 model is the most general of the 3 forms of random effect model that we have fitted. Task 4 model c is the best fitting model and a 95% confidence interval on `corr` does not include 1. The `scale1` and `scale2` parameters of Task 4 model c, are very similar. The significance of `base` in Task 4 model c, is lower than it is in Task 2 and 3.

The fact that `base` is significant in Task 4 model c, suggests that we have not been able to fully account for the initial conditions in this data. Perhaps higher order effects are present. We also suspect that there may be selection effects on the number of patents applied for, as there are very few firms with zero patents at all years in the data, if so its likley that there will be a correlation between the included and random effects.

28.2 Batch Script: patents.do

```
log using patents_s.log, replace
set more off
use patents
#delimit ;
sabre, data 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;
sabre 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, read;
#delimit cr
sabre, case cusip
sabre, yvar pat
sabre, rvar r
sabre, family p
sabre, constant cons
sabre, trans r1_logr r1 * logr
sabre, trans r1_logk r1 * logk
sabre, trans r1_scisect r1 * scisect
sabre, trans r2_logr r2 * logr
sabre, trans r2_logk r2 * logk
sabre, trans r2_scisect r2 * scisect
sabre, trans r2_year3 r2 * year3
sabre, trans r2_year4 r2 * year4
sabre, trans r2_year5 r2 * year5
sabre, trans r2_pat1 r2 * pat1
sabre, trans r2_base r2 * base
sabre, quad a
sabre, mass 36
#delimit ;
sabre, fit r1 r1_logr r1_logk r1_scisect
           r2 r2_logr r2_logk r2_scisect r2_year3 r2_year4 r2_year5;
#delimit cr
sabre, dis m
sabre, dis e
#delimit ;
sabre, fit r1 r1_logr r1_logk r1_scisect
           r2 r2_pat1 r2_logr r2_logk r2_scisect r2_year3 r2_year4 r2_year5;
#delimit cr
sabre, dis m
sabre, dis e
#delimit ;
sabre, fit r1 r1_logr r1_logk r1_scisect
           r2 r2_pat1 r2_base r2_logr r2_logk r2_scisect r2_year3 r2_year4
           r2_year5;
#delimit cr
sabre, dis m
sabre, dis e
sabre, depend y
sabre, mass 24
sabre, ari a
sabre, nvar 4
#delimit ;
sabre, fit r1 r1_logr r1_logk r1_scisect
           r2 r2_logr r2_logk r2_scisect r2_year3 r2_year4 r2_year5;
#delimit cr
sabre, dis m
sabre, dis e
sabre, nvar 4
#delimit ;
sabre, fit r1 r1_logr r1_logk r1_scisect
```

```
          r2 r2_pat1 r2_logr r2_logk r2_scisect r2_year3 r2_year4 r2_year5;
#delimit cr
sabre, dis m
sabre, dis e
sabre, nvar 4
#delimit ;
sabre, fit r1 r1_logr r1_logk r1_scisect
          r2 r2_pat1 r2_base r2_logr r2_logk r2_scisect r2_year3 r2_year4
          r2_year5;
#delimit cr
sabre, dis m
sabre, dis e
sabre, depend n
sabre, model b
sabre, family first=p second=p
sabre, constant first=r1 second=r2
sabre, mass first=36 second=36
sabre, nvar 4
#delimit ;
sabre, fit r1 r1_logr r1_logk r1_scisect
          r2 r2_logr r2_logk r2_scisect r2_year3 r2_year4 r2_year5;
#delimit cr
sabre, dis m
sabre, dis e
sabre, nvar 4
#delimit ;
sabre, fit r1 r1_logr r1_logk r1_scisect
          r2 r2_pat1 r2_logr r2_logk r2_scisect r2_year3 r2_year4 r2_year5;
#delimit cr
sabre, dis m
sabre, dis e
sabre, nvar 4
#delimit ;
sabre, fit r1 r1_logr r1_logk r1_scisect
          r2 r2_pat1 r2_base r2_logr r2_logk r2_scisect r2_year3 r2_year4
          r2_year5;
#delimit cr
sabre, dis m
sabre, dis e
log close
clear
exit
```

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

29.1 Relevant Results from `jtrain_s.log` and Discussion

Task 1. Estimate a linear model for the response `lscrap`, with covariates `grant`, `d89`, `d88` and `grant_1`. Re-estimate the model using the fixed firm effects (`fcode`). What is the main difference between the results from the alternative estimators?

Result/Discussion

Homogeneous linear model

```
Log likelihood =     -292.16964     on     156 residual degrees of freedom

Parameter              Estimate        Std. Err.
---------------------------------------------------
cons                    0.59743         0.20306
d88                    -0.23937         0.31086
d89                    -0.49652         0.33793
grant                   0.20002         0.33828
grant_1                 0.48936E-01     0.43607
sigma                    1.4922
```

Fixed effects model

```
Parameter              Estimate        Std. Err.
---------------------------------------------------
d88                    -0.80216E-01     0.11001
d89                    -0.24720         0.13386
grant                  -0.25231         0.15136
grant_1                -0.42159         0.21122
sigma                   0.50015
```

None of the estimated covariate parameters are significant in the homogenous linear model. In the fixed effects model, both the estimated parameters for `grant` and `grant_1` are negative, and that for `grant_1` is significant, with z statistic -0.42159/0.21122= -1.996. The fixed effects model suggests that firms receiving a training grant have lower scrap rates the following year than those that do not, perhaps this is indicating improved productivity. The problem with this interpretation is that `grant` and `grant_1` are not randomly allocated as firms have chosen whether or not to apply for grants and. not all firms applied.

The coefficient on `d89` is of marginal significance. The value of `sigma` is much smaller in the fixed effects model. The fact that the estimates from the homogenous and fixed effects models are different, suggests that incidental parameters are present.

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

Result/Discussion

Homogeneous linear model

```
Log likelihood =     -292.17613     on     157 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
cons                    0.59743         0.20243
d88                    -0.23641         0.30877
d89                    -0.47775         0.29268
grant                   0.19161         0.32884
sigma                    1.4875
```

Fixed effects model

```
Parameter              Estimate         Std. Err.
___________________________________________________
d88                    -0.14007         0.10735
d89                    -0.42704         0.10041
grant                  -0.82214E-01     0.12687
sigma                   0.50728
```

None of the estimated covariate parameters are significant in the homogenous linear model. In the fixed effects model the estimated parameter for `d89` is very significant. The fixed effects model is suggesting that firms reduced their scrap rates in 1989, but that `grant` had no effect. The value of `sigma` is much smaller in the fixed effects model.

The log likelihood of the homogeneous model of Task 1 is -292.16964 and log likelihood of the homogeneous model of Task 2 is -292.17613. The change in log likelihood is -2(-292.17613+292.16964)= 0.012 98. The sampling distribution of this test statistic is chi-square with 1 df. Under the null hypothesis `grant_1`=0. The test statistic is clearly not significant, suggesting that `grant_1`=0. The same inference is made by the z statistic for `grant_1`. The fact that the estimates from the homogeneous and fixed effects models are different, suggests that incidental parameters are present. There is no log likelihood that we can use to compare models for the fixed effects estimator.

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

Result/Discussion

My preferred model is the fixed effects model of Task 1. The negative estimated parameter on d89, suggests that 1989 had lower scrap rates than either 1987 or 1988.

Task 4. Re-estimate the fixed effects models of Tasks 1 and 2 using adaptive quadrature and mass 12. Compare the fixed and random effect model inferences. What do you find?

Result/Discussion

Random effects model with `grant_1`.

```
Log likelihood =     -201.25249     on     155 residual degrees of freedom

Parameter              Estimate        Std. Err.
---------------------------------------------------
cons                    0.59743        0.20118
d88                    -0.93319E-01    0.10701
d89                    -0.27095        0.12916
grant                  -0.21507        0.14515
grant_1                -0.37369        0.20165
sigma                   0.48861        0.33268E-01
scale                    1.3953        0.14000
```

Random effects model without `grant_1`.

```
Log likelihood =     -202.93415     on     156 residual degrees of freedom

Parameter              Estimate        Std. Err.
---------------------------------------------------
cons                    0.59743        0.20031
d88                    -0.14510        0.10525
d89                    -0.42969        0.98513E-01
grant                  -0.67913E-01    0.12384
sigma                   0.49783        0.33877E-01
scale                    1.3852        0.13912
```

The log likelihood of the random effects model with `grant_1` is -201.25249 and log likelihood of the model without is -202.93415. The change in log likelihood is -2(-202.93415+201.25249)= 3.3633. The sampling distribution of this test statistic is chi-square with 1 df, and suggests that `grant_1` is not significant, the z statistic for `grant_1` gives a similar result. It may be worth estimating a model without `grant` but with `grant_1`.

Both the models of Task 4 are significant improvements over their respective homogenous versions (Task 1 and 2), suggesting that random effects are present. The differences between the parameter estimates of the fixed effect and random effect version of the same model suggests that the assumption of independence between the random effects and the included covariates may not hold, but a random effects analysis in which the time averages of the covariates are significant would be needed to confirm this.

29.2 Batch Script: `jtrain.do`

```
log using jtrain_s.log, replace
set more off
use jtrain
#delimit ;
sabre, data year fcode employ sales avgsal scrap rework tothrs union grant
            d89 d88 totrain hrsemp lscrap lemploy lsales lrework lhrsemp
            lscrap_1 grant_1 clscrap cgrant clemploy clsales lavgsal
            clavgsal cgrant_1 chrsemp clhrsemp;
sabre year fcode employ sales avgsal scrap rework tothrs union grant d89 d88
      totrain hrsemp lscrap lemploy lsales lrework lhrsemp lscrap_1 grant_1
      clscrap cgrant clemploy clsales lavgsal clavgsal cgrant_1 chrsemp
      clhrsemp, read;
#delimit cr
sabre, case fcode
sabre, yvar lscrap
sabre, fam g
sabre, constant cons
sabre, lfit d88 d89 grant grant_1 cons
sabre, dis m
sabre, dis e
sabre, fefit d88 d89 grant grant_1
sabre, dis m
sabre, dis e
sabre, lfit d88 d89 grant cons
sabre, dis m
sabre, dis e
sabre, fefit d88 d89 grant
sabre, dis m
sabre, dis e
sabre, quad a
sabre, mass 12
sabre, fit d88 d89 grant grant_1 cons
sabre, dis m
sabre, dis e
sabre, fit d88 d89 grant cons
sabre, dis m
sabre, dis e
log close
clear
exit
```

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

30.1 Relevant Results from `wagepan2_s.log` and Discussion

Task 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.

Result/Discussion

- sabre, trans edd81 educ * d81
- sabre, trans edd82 educ * d82
- sabre, trans edd83 educ * d83
- sabre, trans edd84 educ * d84
- sabre, trans edd85 educ * d85
- sabre, trans edd86 educ * d86
- sabre, trans edd87 educ * d87

Task 2. Estimate a linear model for the response `lwage` with the covariates `espersq, union, married, d81-d87, edd81-edd97`. Re-estimate the model using the respondent fixed effects (`nr`). What is the main difference between the results from the alternative estimators?

Result/Discussion

Homogeneous linear model

```
Log likelihood =     -3023.3871     on    4341 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
cons                     1.3126         0.21684E-01
expersq                 0.10610E-02     0.33426E-03
union                   0.17733         0.17140E-01
married                 0.12840         0.15590E-01
d81                    -0.81625         0.14562
d82                    -0.82033         0.14716
d83                    -0.83814         0.14920
d84                    -0.80049         0.15190
```

d85	-0.84403	0.15531
d86	-0.85702	0.15944
d87	-0.88431	0.16439
edd81	0.77787E-01	0.12085E-01
edd82	0.81445E-01	0.12158E-01
edd83	0.85194E-01	0.12239E-01
edd84	0.86192E-01	0.12334E-01
edd85	0.92685E-01	0.12443E-01
edd86	0.97193E-01	0.12553E-01
edd87	0.10227	0.12675E-01
sigma	0.48508	

Fixed effects model

Parameter	Estimate	Std. Err.
expersq	-0.60437E-02	0.86338E-03
union	0.78976E-01	0.19335E-01
married	0.47434E-01	0.18330E-01
d81	0.98420E-01	0.14602
d82	0.24720	0.14940
d83	0.40881	0.15574
d84	0.63992	0.16526
d85	0.77294	0.17801
d86	0.96993	0.19420
d87	1.1888	0.21361
edd81	0.49906E-02	0.12224E-01
edd82	0.16510E-02	0.12332E-01
edd83	-0.26621E-02	0.12511E-01
edd84	-0.98257E-02	0.12761E-01
edd85	-0.92145E-02	0.13074E-01
edd86	-0.12138E-01	0.13444E-01
edd87	-0.15789E-01	0.13870E-01
sigma	0.35119	

Most of the estimated covariate parameters are significant in the homogenous linear model. The fixed effects covariate parameter model estimates are very different to those of the homogeneous linear model, also non of the interaction effects of `educ` with year are significant in the fixed effects model.

The value of `sigma` is smaller in the fixed effects model. The fact that the estimates from the homogenous and fixed effects models are different, suggests that incidental parameters are present.

Task 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?

Result/Discussion

Homogeneous linear model

```
Log likelihood =     -3149.2321     on    4348 residual degrees of freedom
```

Parameter	Estimate	Std. Err.
cons	1.3454	0.22199E-01
expersq	-0.20775E-02	0.27670E-03
union	0.17680	0.17624E-01
married	0.15213	0.15943E-01
d81	0.11869	0.30320E-01
d82	0.18434	0.30638E-01
d83	0.24312	0.31337E-01
d84	0.33215	0.32448E-01
d85	0.41121	0.34132E-01
d86	0.50387	0.36497E-01
d87	0.59522	0.39612E-01
sigma	0.49889	

Fixed effects model

Parameter	Estimate	Std. Err.
expersq	-0.51855E-02	0.70453E-03
union	0.80002E-01	0.19313E-01
married	0.46680E-01	0.18313E-01
d81	0.15119	0.21952E-01
d82	0.25297	0.24422E-01
d83	0.35444	0.29246E-01
d84	0.49011	0.36231E-01
d85	0.61748	0.45249E-01
d86	0.76550	0.56135E-01
d87	0.92502	0.68782E-01
sigma	0.35104	

The log likelihood of the homogeneous model of Task 1 is -3023.3871 and log likelihood of the homogeneous model of Task 2 is -3149.2321. The change in log likelihood is -2(-3149.2321+3023.3871)= 251. 69. The sampling distribution of this test statistic is chi-square with 7 df. Under the null hypothesis the interaction effects of educ with `year` take the value 0. The test statistic is clearly significant, suggesting that these interaction effects are present in the model However, this inference is not supported by the fixed effect model of Task 2.

The fact that the estimates from the homogenous and fixed effects models of Task 3 are different, suggests that incidental parameters are present. The common covariate parameter estimates from the fixed effect model from Task 2 and Task 3 are very similar, and the fixed effects model of Task 3 is more parsimonious. There is no log likelihood that we can use to compare models for the fixed effects estimator.

Task 4. Re-estimate the fixed effects model of Task 2 using adaptive quadrature with mass 12. Compare the fixed and random effect model inferences. What do you find?

Result/Discussion

```
Log likelihood =     -2943.6408     on    4341 residual degrees of freedom

Parameter              Estimate         Std. Err.
___________________________________________________
expersq              -0.22011E-02      0.84074E-03
union                 0.11906          0.19420E-01
married               0.77160E-01      0.18343E-01
d81                   0.76501E-01      0.14475
d82                   0.14187          0.14836
d83                   0.20791          0.15482
d84                   0.33558          0.16434
d85                   0.36988          0.17689
d86                   0.45310          0.19274
d87                   0.54148          0.21175
edd81                 0.11842E-01      0.12115E-01
edd82                 0.12442E-01      0.12244E-01
edd83                 0.12459E-01      0.12441E-01
edd84                 0.96787E-02      0.12704E-01
edd85                 0.13763E-01      0.13023E-01
edd86                 0.14902E-01      0.13398E-01
edd87                 0.15852E-01      0.13828E-01
sigma                 0.35294          0.41047E-02
scale                  1.3442          0.45440E-01
```

The log likelihood of the random effects model is -2943.6408 and log likelihood of the homogeneous model is -3023.3871. The change in log likelihood is -2(-3023.3871+2943.6408)= 159. 49. The sampling distribution of this test statistic is not chi-square with 1 df. Under the null hypothesis `scale` has the value 0, it can only take values >0 under the alternative. The correct p value for this test statistics is obtained by dividing the naive p value of 159. 49 for 1 degree of freedom by 1/2, and so its clearly significant.

There are some differences between the parameter estimates of the fixed effect and random effect versions of the same model, but these differences are not large, e.g. both models find no evidence for an interaction between `educ` and year. Perhaps the assumption of independence between the random effects and the included covariates holds, but a random effects analysis in which the time averages of the covariates are non significant would be needed to confirm this.

30.2 Batch Script: `wagepan2.do`

```
log using wagepan2_s.log, replace
set more off
```

```
use wagepan2
#delimit ;
sabre, data 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;
sabre 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,
      read;
#delimit cr
sabre, case nr
sabre, yvar lwage
sabre, family g
sabre, constant cons
sabre, trans edd81 educ * d81
sabre, trans edd82 educ * d82
sabre, trans edd83 educ * d83
sabre, trans edd84 educ * d84
sabre, trans edd85 educ * d85
sabre, trans edd86 educ * d86
sabre, trans edd87 educ * d87
#delimit ;
sabre, lfit expersq union married d81 d82 d83 d84 d85 d86 d87 edd81 edd82
            edd83 edd84 edd85 edd86 edd87 cons;
#delimit cr
sabre, dis m
sabre, dis e
#delimit ;
sabre, fefit expersq union married d81 d82 d83 d84 d85 d86 d87 edd81 edd82
             edd83 edd84 edd85 edd86 edd87;
#delimit cr
sabre, dis m
sabre, dis e
sabre, lfit expersq union married d81 d82 d83 d84 d85 d86 d87 cons
sabre, dis m
sabre, dis e
sabre, fefit expersq union married d81 d82 d83 d84 d85 d86 d87
sabre, dis m
sabre, dis e
sabre, fefit expersq married union d81 d82 d83 d84 d85 d86 d87
sabre, dis m
sabre, dis e
sabre, quad a
sabre, mass 12
#delimit ;
sabre, fit expersq union married d81 d82 d83 d84 d85 d86 d87 edd81 edd82
           edd83 edd84 edd85 edd86 edd87;
#delimit cr
sabre, dis m
sabre, dis e
log close
clear
exit
```