Multi-level Modelling Assignment

Student number: 8520000

Part A

Question 1

Provide some summary statistics about the levels in the data. How many units are there at each level (overall N of each level), and how many units are there within levels (N within each level).

There are 12 different Government office regions (level 3), 4429 different households (level 2), and 4655 individuals (level 1).The table below provides information for every region on the number of households, individuals, and the mean number of individuals per household.

Region	Number of households	Number of individuals	Average number of individuals per household
1	142	148	1.042253
2	483	506	1.047619
3	333	353	1.060060
4	333	356	1.069069
5	309	330	1.067961
6	350	364	1.040000
7	531	554	1.043314
8	512	542	1.058594
9	360	374	1.038889
10	333	353	1.060060
11	408	425	1.041667
12	335	350	1.044776

However there were some missing values, after dropping rows with incomplete values the dataset contains 12 different Government office regions (level 3), 3883 different households (level 2), and 4064 individuals (level 1).The table below provides information for every region on the number of households and individuals, and the mean number of individuals per household. The rest of this assignment will be conducted with the missing values removed.

Region	Number of households	Number of individuals	Average number of individuals per household
1	123	128	1.040650
2	435	454	1.043678
3	291	304	1.044674
4	278	297	1.068345
5	262	279	1.064885
6	321	333	1.037383
7	461	478	1.036876
8	458	483	1.054585
9	329	342	1.039514
10	279	295	1.057348
11	369	384	1.040650
12	277	287	1.036101

Question 2

Specify the generalised equations of a three level multilevel model with random slopes at levels 2 and 3. State the assumptions of the model.

The generalised equation is:

$Y_{ijk} = \beta_0 + \beta_1 x_{1ijk} + v_{0k} + v_{1k} x_{1ijk} + u_{0jk} + u_{1ijk} x_{1ijk} + e_{ijk}$

Where Y_{ijk} is the expected value of our dependant variable for individual i in household j in region k, x_{1ijk} is a independant variable with random coeffecients at levels two and three, β_0 is the mean intercept over all slopes, β_1 is the average slope across all groups (the average in y change across all groups for a 1 unit change in x_{1ijk}), and v_{0k}, $v_{1k}x_{1ijk}$, u_{0jk}, $u_{1ijk}x_{1ijk}$, e_{ijk} are the random effects associated with level two, level three, and the individual level respectively.

v_k is the effect associated with region k, u_{jk} is the effect associated with household j within region k, and e_{ijk} is the residual error term (the difference between the mean value in household j, and the value for individual i).

Question 3

Start from the single level null model and add in the household and then region levels.

A. Display the model coefficients in a table.

	Dependent variable:Neighborhood Mistrust			
	Single level null	Two levels (household)	Two levels (region)	Three levels
Constant	2.385^{***}	2.382^{***}	2.383^{***}	2.379^{***}
	(2.365, 2.406)	(2.362, 2.403)	(2.343, 2.422)	(2.339, 2.419)

B. Does the addition of the household level improve the fit of the model?

We can use the likelihood-ratio test to asses whether the 2 level household model fits the data better than the single level null model. To do so we calculate the likelihood ratio test statistic, and compare the result to a chi-squared distribution on one degree of freedom.

The likelihood ratio test statistic for these models is approximately 94.45, and the 5% point on a chi-squared distribution with two degrees of freedom is approximately 3.84.

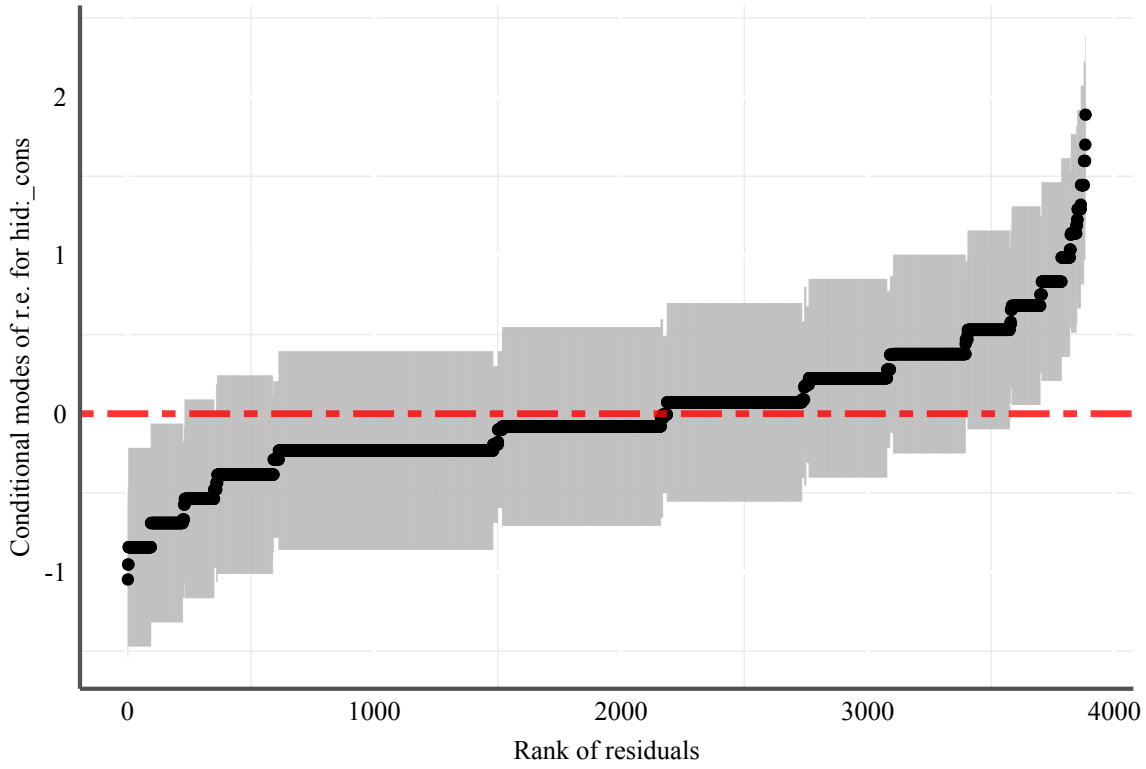
We therefore have strong reasons to believe that the addition of the household level improves the fit of the model.

C.What evidence is there for an improvement in fit?

There are a number of reasons to believe the addition of the household level improves the fit of the model.

First, the likelihood-ratio test conducted in the answer to the previous question provides strong evidence that we have a statistically significant improvement in model fit.

We can also create a 'caterpillar' plot using the ggplot2 package which displays ranked residuals from the household model (in black), with their 95% confidence intervals (in grey), as well as a red line for the average effect.



We can see that the confidence individuals of a large number of households do not overlap with this line. In a single level model, all these households would be fit in a similar way, however there is clearly large differences between households, therefore we can expect an improvement in fit in moving to a two level model with households at level 2.

We can also conduct likelihood-ratio tests to asses whether the addition of region at level two provides an improvement in fit over a single level model, and whether a three-level model with household at level two and region at level three provides an improvement in fit over the two-level models.

The likelihood-ratio test statistic for the addition of region at level two is approximately 20.19, which suggest a highly statistically significant improvement in fit.

The likelihood-ratio test statistic for the three-level model is approximately 94.3 when compared to the two-level model with region at level two, and approximately 20 when compared to the two-level model with household at level two. It both cases the test suggests strongly statistically significant improvements in model fit.