Multi-level Modelling Assignment

Student number: 8520000

This assignment was fulfilled in R (R Core Team, 2013), all graphs were created using base R, the ggplot2 package (Wickham, 2016) or the lattice package (Sarkar, 2008), all tables and equations were created using the TinyTex distribution for LaTex (Xie, 2020) or sjPlot (Lüdecke D, 2020), and all multi-level models were fitted using the lme4 package (Bates, et al. 2015).

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 data set 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 generalized equations of a three level multilevel model with random slopes at levels 2 and 3. State the assumptions of the model.

The generalized 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 dependent variable for individual i in household j in region k, x_{1ijk} is a independent variable with random coefficients 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).

The model assumes that residuals are normally distributed at each level, that there is no heteroscedasticity at each level, and that variance of the residuals is equal within all groups, but not between all groups. The effect of x_{1ijk} varies between groups.

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)

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

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.

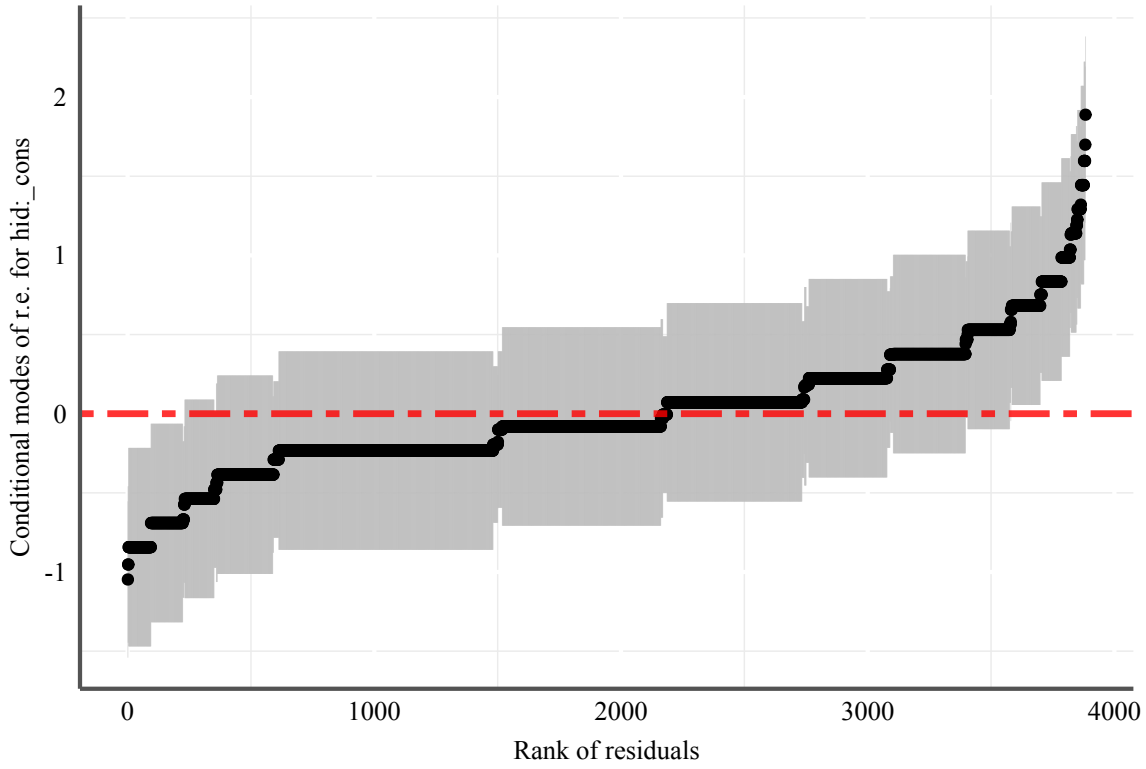
We therefore have strong reasons to believe that the addition of the household level improves the fit of the model.

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.

Question 4 - Calculate the VPC at the household and regional levels for the 3 level variance components model.

The residual variances for our model are 0.1672, 0.2588, and 0.0036 for the individual, household, and region levels respectively.

The region VPC is calculated as the ratio of the region variance to the overall variance of the model, therefore:

Region VPC $\approx$ 0.0036 / (0.1672 + 0.2588 + 0.0036) $\approx$ 0.008

The household VPC is calculated as the ratio of the household variance to the overall variance of the model, therefore:

Household VPC $\approx$ 0.2588 / (0.1672 + 0.2588 + 0.0036) $\approx$ 0.602

We can therefore say that approximately 0.8% of the variance in our model is between regions, and approximately 60% is within regions, between households (with the remaining variance ($\approx$ 39%) happening within households, between individuals).

Question 5 - Add in the following explanatory variables in the model (random intercepts model only; no random slopes or coefficients)- age, sclfsato, urban, female, hhtenure, hiqual3. Take out any non-significant associations.

The variable hhtenure was re-coded as two binary dummy variables: 'local authority/housing association rent' corresponds to the variable 'rent_local_auth' and 'private rent' corresponds to the variable 'rent_private'. 'Owner/mortgaged' is the reference category.

The variable hiqual3 was also re-coded as binary dummy variables: 'school level qualification' corresponds to the variable 'school_qual', and 'No qualification' corresponds the the variable 'no_qual'. 'Degree or equivalent ' is the reference category.

The minimum value for age is 16. For ease of interpretation, 16 was subtracted from the age variable.

Variables were added to the model one by one (dummies of a same variable were added at the same time), each time a likelihood-ratio test was conducted to asses whether the addition of the variable was statistically significant. Female was the only variable which did not pass the test (LR test-statistic of 0.230), and was therefore dropped.

A - Display the model coefficients in a table.

	Dependent variable: Neighborhood Mistrust
	Random Intercept Model
Constant	2.563^{***}
	(2.471, 2.656)
age	-0.005^{***}
	$(-0.006, -0.004)$
sclfsato	-0.059^{***}
	$(-0.072, -0.047)$
urban	0.237^{***}
	(0.191, 0.284)
school_qual	0.074^{***}
	(0.033, 0.116)
no_qual	0.006
	$(-0.060, 0.072)$
rent_local_auth	0.234^{***}
	(0.179, 0.289)
rent_private	0.122^{***}
	(0.059, 0.186)

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

B. Interpret the coefficients.

The intercept is ≈ 2.563, this is the grand mean of our model, and the predicted value of neighborhood mistrust when all other variables are set to zero.

The coefficient for age is ≈ -0.005, earlier we subtracted the minimum value of age (16) from all values of age, therefore, all other variables being fixed, we expect in decrease of approximately 0.005 in neighborhood mistrust for every additional year of age.

The coefficient for sclfsato is ≈ -0.059, this variable is not centered in any way, we therefore expect the predicted value for neighborhood mistrust to decrease by 0.059 for an unit increase in sclfsato.

The four remaining variables (urban, school_qual, no_qual, rent_local_auth, rent_private) are binary, their coefficeints therefore correspond to the expected increase in neighborhood mistrust when their respective values are 1 (all other variables being fixed).

Question 6 - Add in random slopes for age, at the household and regional level

A. Does this improve the fit of the model?

We can use the likelihood-ratio test to asses whether the model with age as random slopes at levels two and three fits the data better than the random intercepts model.

The test statistic is approximately 30.05.The 95% critical value on a chi-squared distribution with four degrees of freedom is 9.48. We therefore have strong evidence of an improvement in fit.

B. Take out from the model any non-significant random slope(s) for age

In the above question we established that adding random slopes for age at level two and level three lead to a statistically significant improvement in fit.

However when we add the random slopes individually we find that the likelihood-ratio test statistic is approximately 30.498 for a model with random slopes for age at the household level and 1.096 for a model with random slopes for age at the household level.

The 95% critical value on a chi-squared distribution on two degrees of freedom is 5.99, so the random slope for age at the region level is not statistically significant and is therefore dropped.

Our model output is therefore:

	Random Slope hid Model
(Intercept)	2.57^{***}
	(0.05)
age	-0.00^{***}
	(0.00)
sclfsato	-0.06^{***}
	(0.01)
urban	0.23^{***}
	(0.02)
school_qual	0.07^{***}
	(0.02)
no_qual	0.01
	(0.03)
rent_local_auth	0.22^{***}
	(0.03)
rent_private	0.12^{***}
	(0.03)
AIC	7532.33
BIC	7614.36
Log Likelihood	−3753.16
Num. obs.	4064
Num. groups: hid	3883
Num. groups: region	12
Var: hid (Intercept)	0.34
Var: hid age	0.00
Cov: hid (Intercept) age	−0.00
Var: region (Intercept)	0.00
Var: Residual	0.15

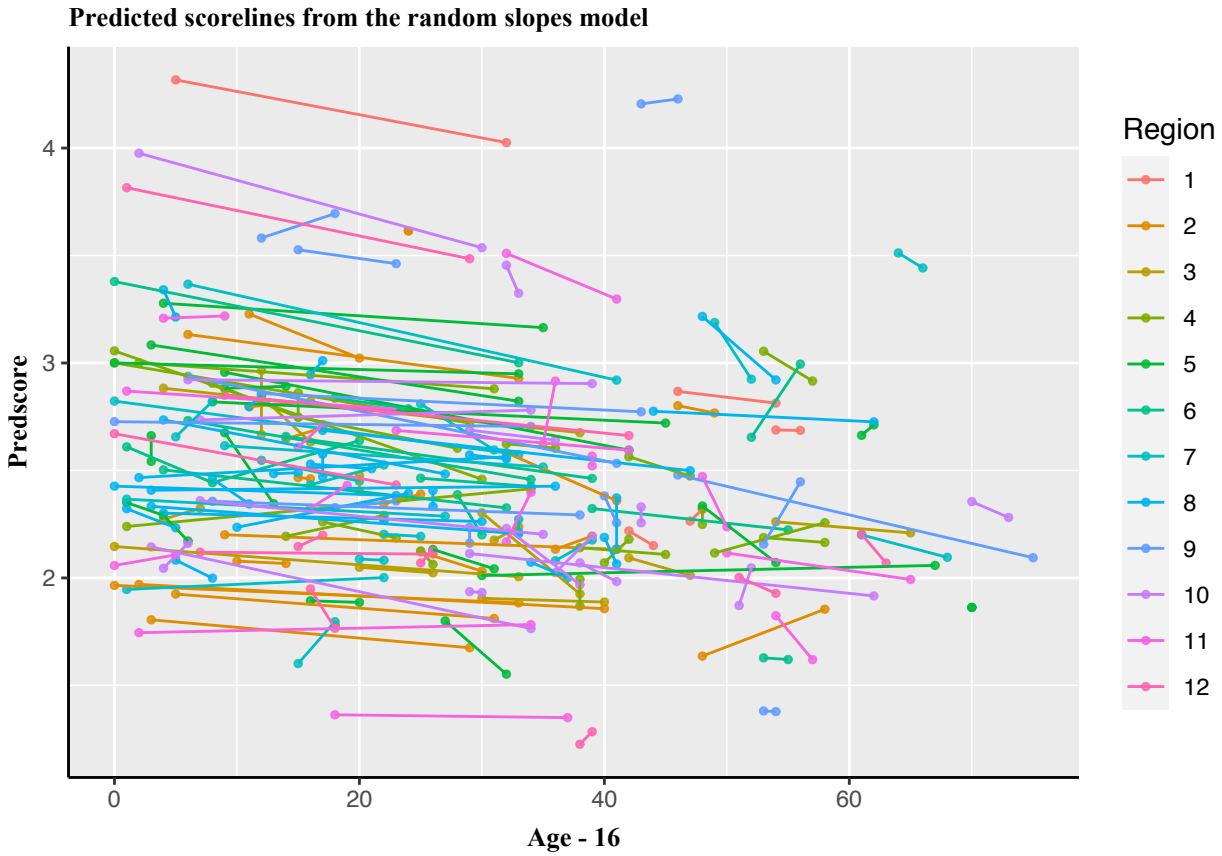
$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

C.Interpret the random slope(s) for age, with the help of graphs.

Due to the reasoning above, our model only contains a single random slope for age (at the household level).

As we can see from the table above, the effect of the random slope for age at the hid level is very small, (the table displays it as 0.00 due to rounding but it's value is approximately 0.00004). However we know from our likelihood ratio test that it is statistically significant.

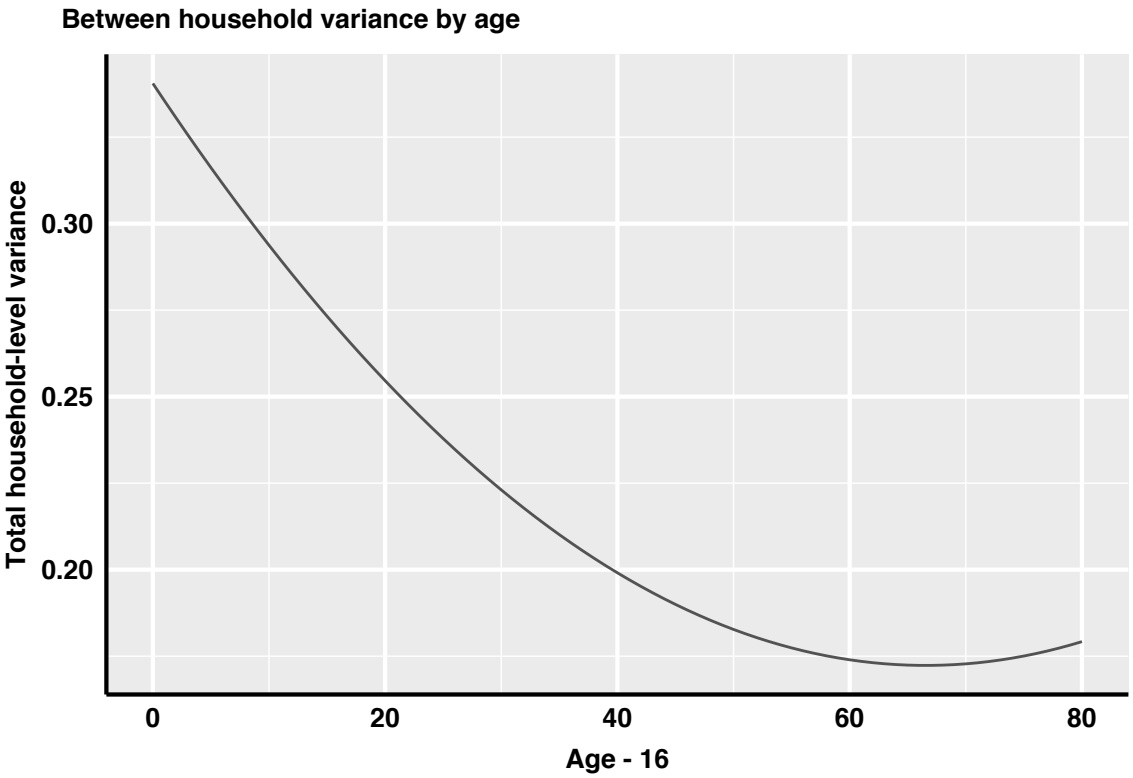
We can create a plot showing the variation between households in how the predicted score for neighborhood mistrust varies according to age. For greater legibility this plot only includes households with at least two members, and the lines have been colored to represent regions.



As expected, the general effect of age is quite small. For individuals in some households there is a large effect of age on the predicted value for neighborhood mistrust, however for some households this effect is much smaller.

The lone dot in the bottom right of the graph is due to both members of household 1456 having the same predicted score (1.862953).

Using ggplot2, we can create a plot showing how between-household variance changes as a function of age.

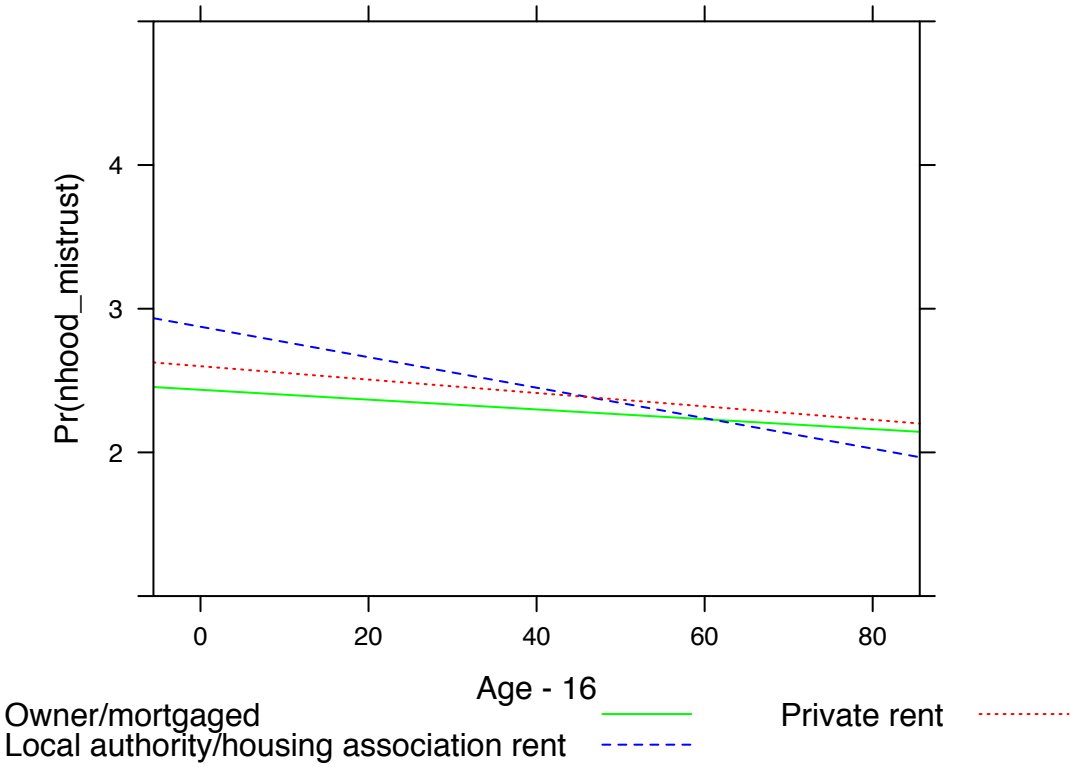


We can see that the variance in neighborhood mistrust between households is much larger for younger households.

Question 7 - Examine potential interaction effects between age and the other explanatory variables- interpret the interactions using graphs. Is the random slope of age explained by these interaction terms?

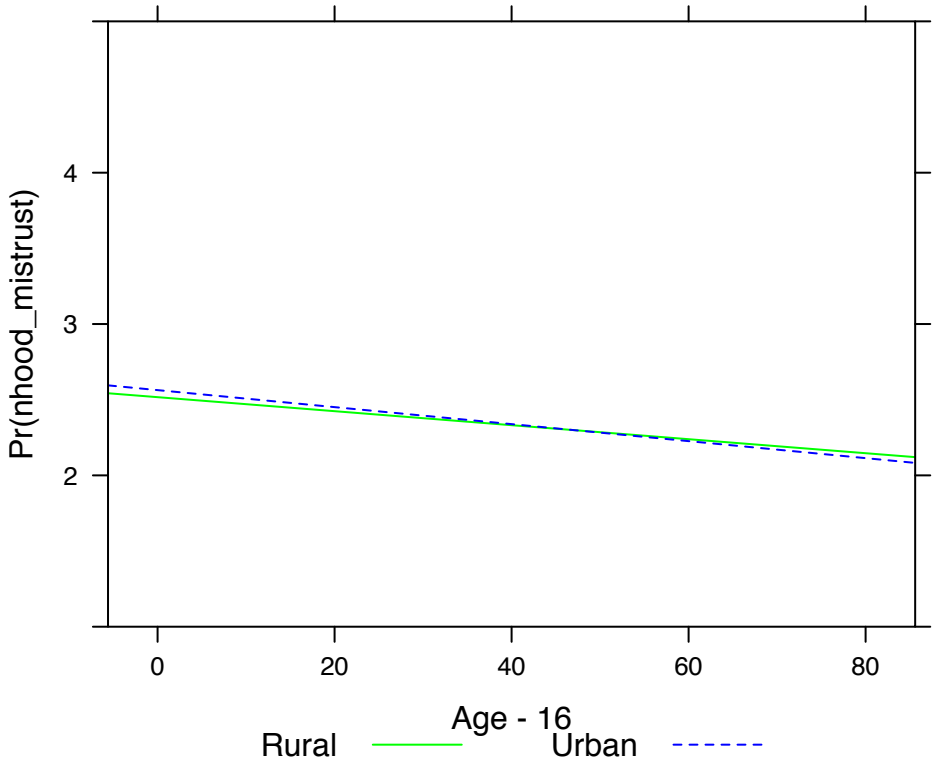
There are six different possible interaction terms for age. Adding each of these in turn to the random intercept model from question 5, and conducting likelihood ratio tests, we find significant improvements in fit from the interaction terms with worry_crime (the test statistic is approximately 47.89), the household tenure dummy variables (the test statistic is approximately 20.57), and urban (the test statistic is approximately 5.36).

We can use graphs to examine these interactions, for example the test conducted above implies that the effect of age on neighborhood mistrust varies according to the type of housing tenure. Using the lattice package:



This illustrates that the negative effect of age on neighborhood mistrust is strongest in households that rent from local authorities. For households who own or have a mortgage on their residence, and households that rent privately, there is a similar effect of age on neighborhood mistrust.

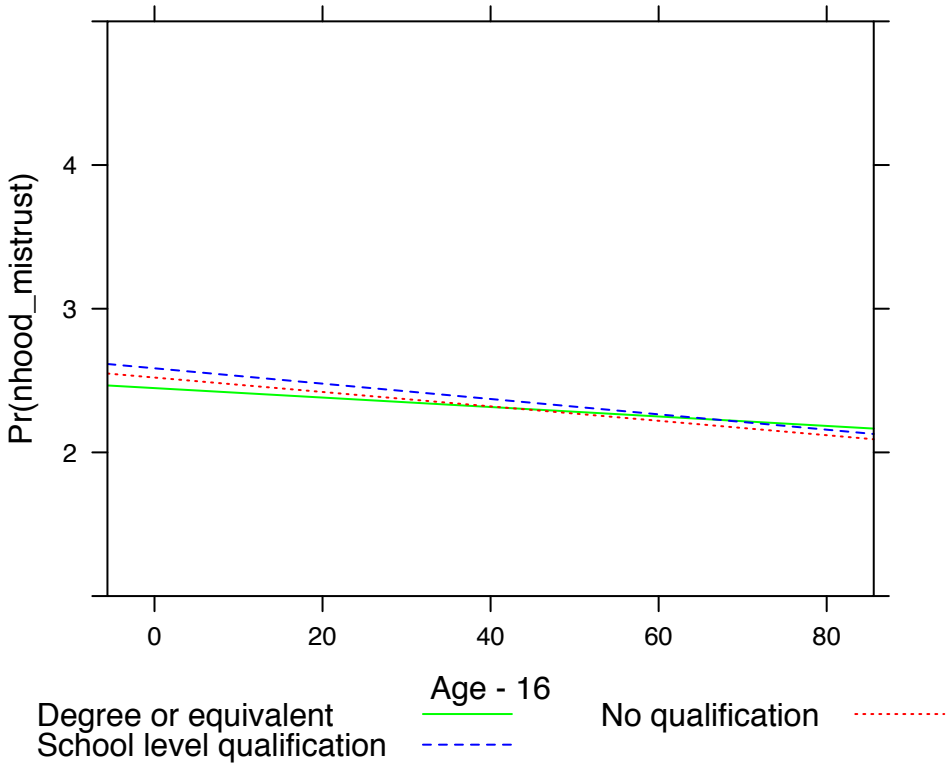
Similarly for the variable 'Urban':



This illustrates the effect of age on neighborhood mistrust varies between urban and rural households. We can see that the effect of age on neighborhood mistrust is stronger in urban areas for younger people, and stronger in rural areas for older people. However the interaction is very weak.

There was no significant improvement in fit from adding interaction terms with female (the test statistic is approximately 0.52), sclfsato (the test statistic is approximately 0.66), and education dummy variables (the test statistic is approximately 2.48).

We can try to visualize these non-significant interactions as above, for example education levels:



It appears that young people with no qualifications or only school-level qualifications have higher levels of neighborhood mistrust than those with degrees, but that this effect is reversed in older people. Whilst it is easy to think of mechanisms through which this may happen, it is important to remember that this interaction is not statistically significant, and we therefore can't rule out that the effect is explained by chance.

It is also important to note that this way of graphing average effect lines can be misleading, for example there are no 16 year-olds with a degree, yet the model will still predict probabilities for them, which can be seen in the lines extending all the way along the x axis.

Comparing a three level model random intercept model with all the significant variables (age, sclfsato, urban, school_qual, no_qual, rent_local_auth, rent_private, ageXworry_crime, ageXurban, ageXrent_local_auth, and ageXrent_private) to a random slope model with the same explanatory variables and with a random slope for age at the household level, we find that the addition of a random slope model provides a statistically significant improvement in fit (the test statistic is approximately 33.62). This suggests that the interaction terms do not fully explain the random slope for age.

Question 8. After fitting all the explanatory variables, is a three level model still appropriate?

Comparing the model with all the significant variables included, a random slope for age at level 2, and variance partitioned at level 3, to a same model without the level three element, we find evidence for a significant improvement in fit when the level three element is included. The test statistic is approximately 4.79, which is significant at the 0.95 level, but not at the 0.99 level. Given that the evidence for an improvement in fit is rather weak, it could be argued that the more parsimonious two-level model is more appropriate.

Question 9. Display and interpret the parameters of your final model in a table.

	Final Model
(Intercept)	2.41***
	(0.06)
age	−0.00
	(0.00)
sclfsato	−0.06***
	(0.01)
urban	0.34***
	(0.05)
school_qual	0.08***
	(0.02)
no_qual	0.02
	(0.03)
rent_local_auth	0.44***
	(0.05)
rent_private	0.17**
	(0.06)
ageXworry_crime	0.00***
	(0.00)
ageXurban	−0.00**
	(0.00)
ageXrent_local_auth	−0.01***
	(0.00)
ageXrent_private	−0.00
	(0.00)
AIC	7465.38
BIC	7566.34
Log Likelihood	−3716.69
Num. obs.	4064
Num. groups: hid	3883
Var: hid (Intercept)	0.39
Var: hid age	0.00
Cov: hid (Intercept) age	−0.00
Var: Residual	0.13

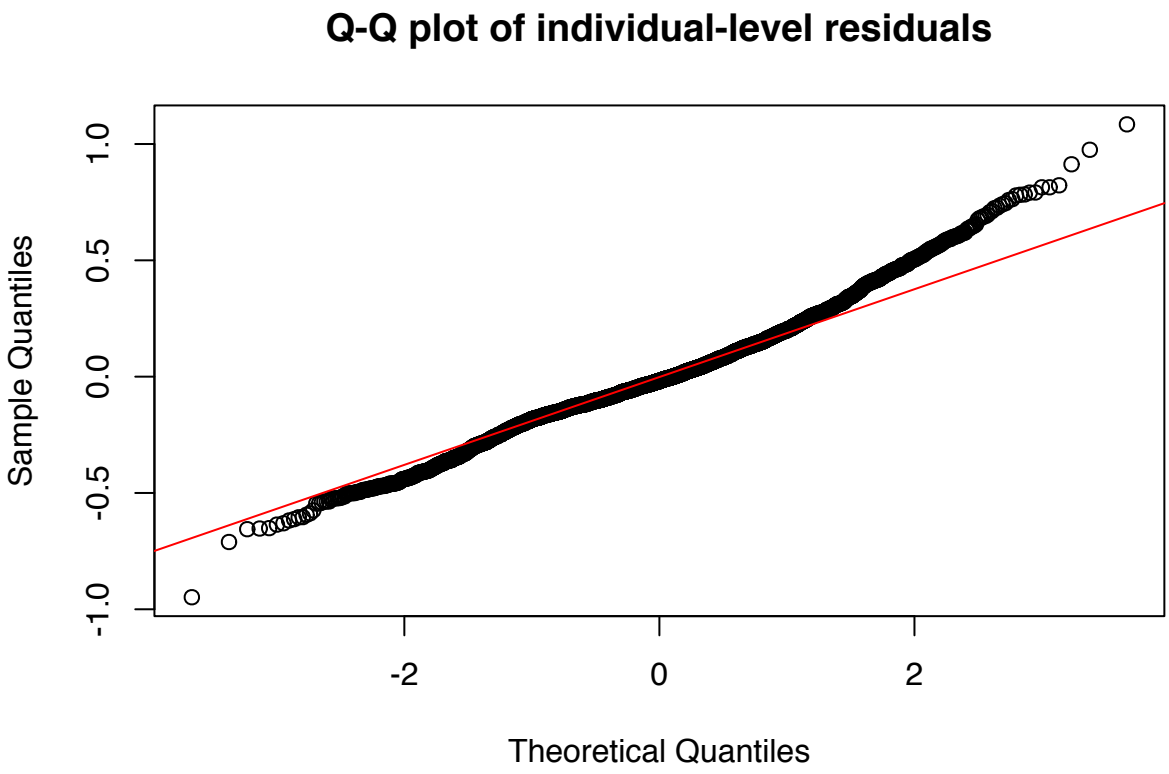
$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Question 10. Examine residuals at all three levels- are the assumptions of the regression model met?

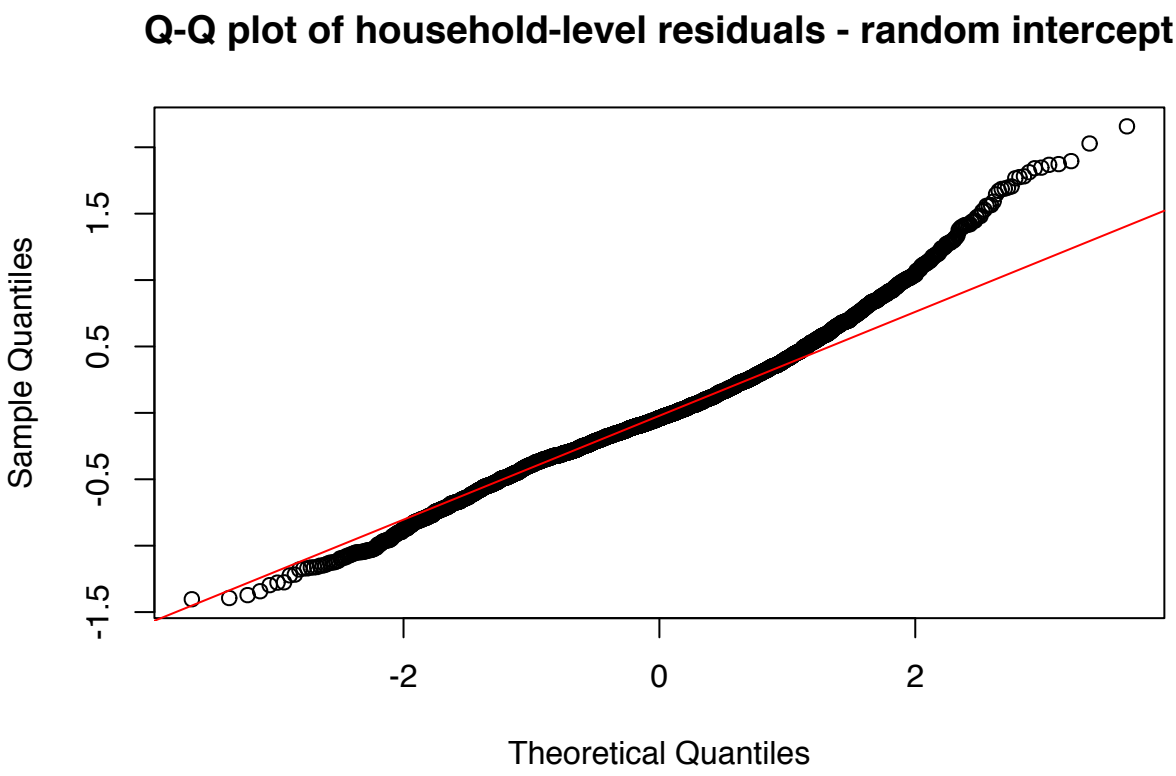
Having decided that a three level model did not provide enough of an improvement in fit over a two level model to justify the loss in parsimony, I will only test the residuals at level one and two.

First we can create 'q-q plots' to test the normality assumption for each main component of our models residuals: the individual level, the household level random intercept, and the household level random slope.

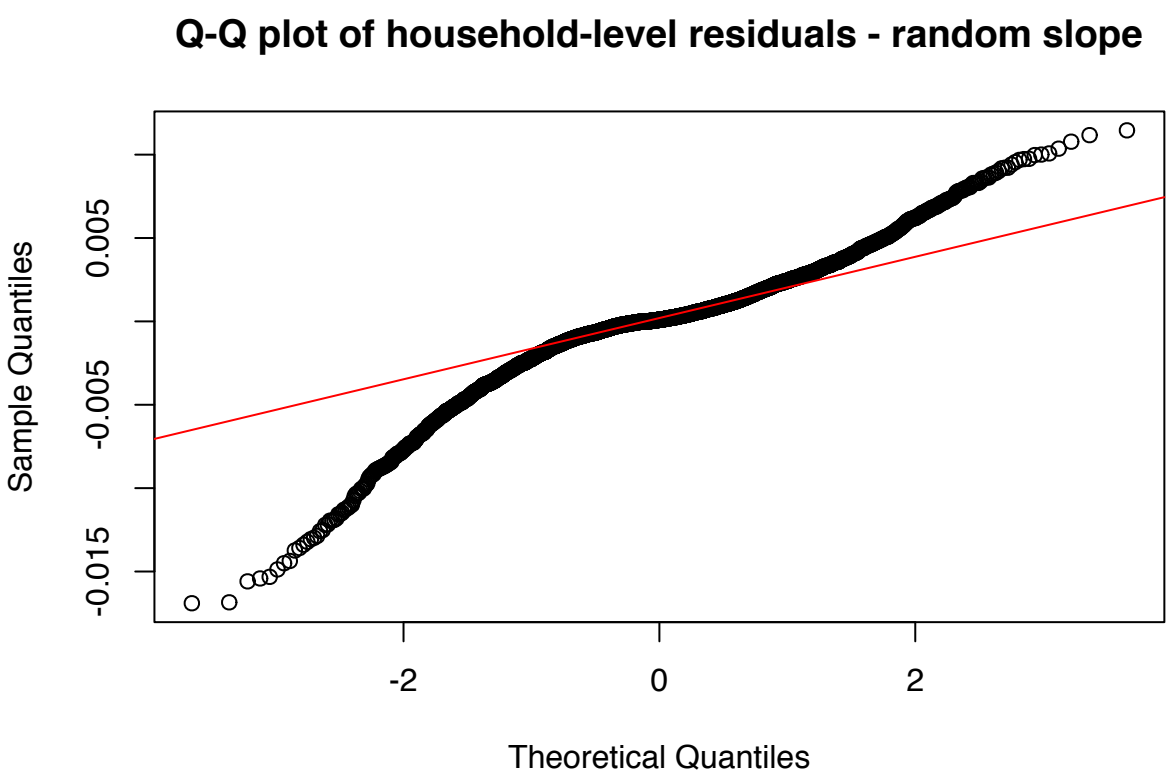
First at the individual level:



Then the household-level residuals, for the random intercept:



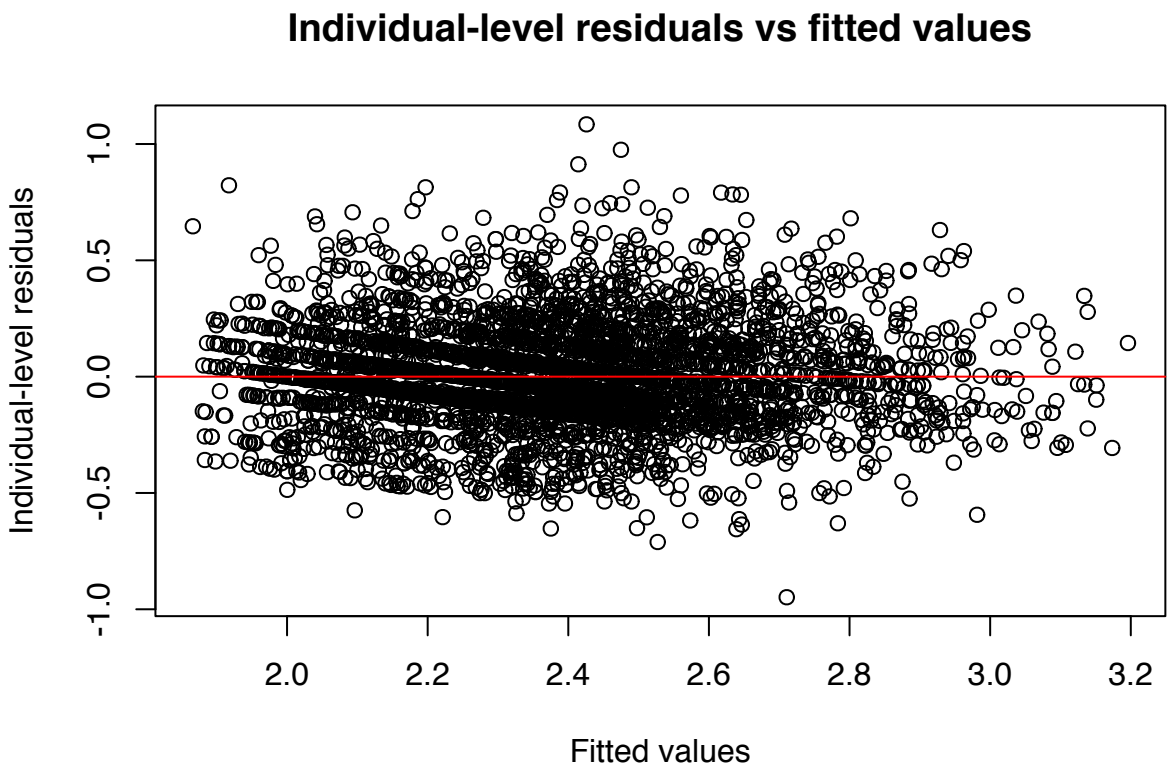
Then the household-level residuals, for the random slope:



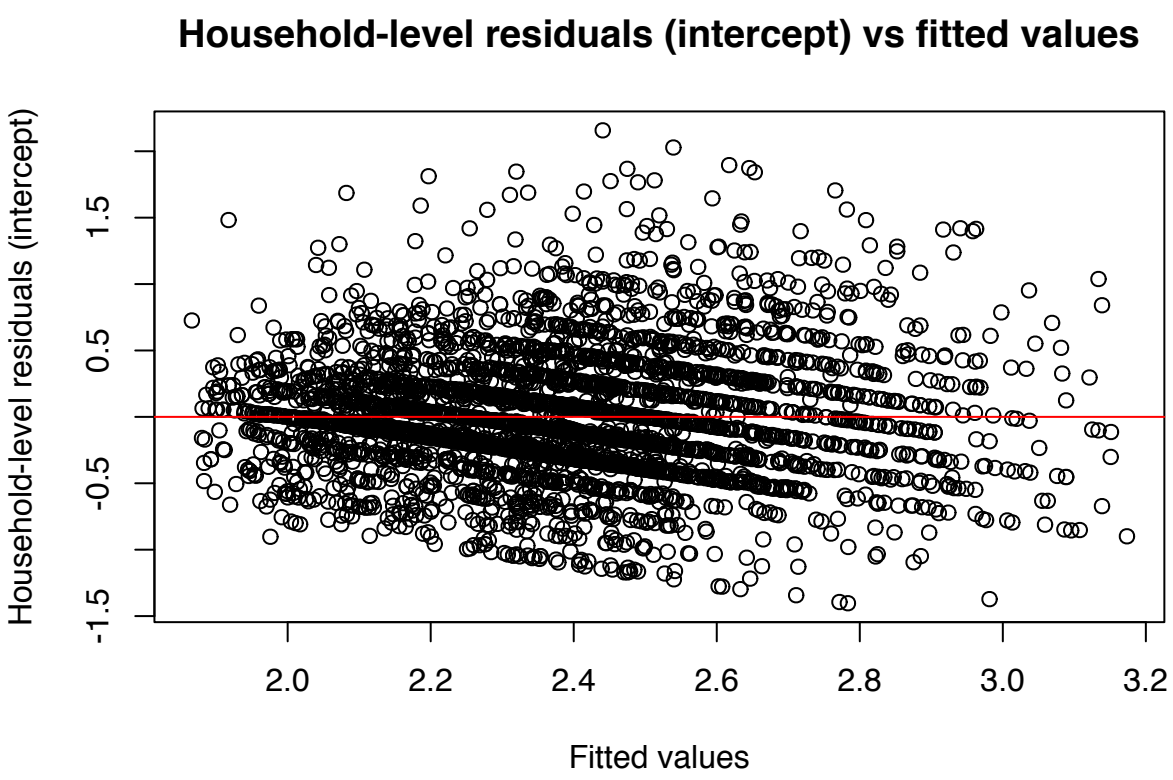
In each case, and in particular, when looking at residuals from the level two slope, we can see that the distribution of our residuals are more 'fat-tailed' than we would expect if the residuals followed a normal distribution. This means that more data takes on values at the extremes of the distribution, and less at the center. The difference is not huge, but we should be wary that the normality assumption is stretched in our model.

We can also check to see if the assumption of homoskedasticity (that residuals are distributed equally over our model's predictions) is met. As above we can look at the individual level, the household level random intercept, and the household level random slope.

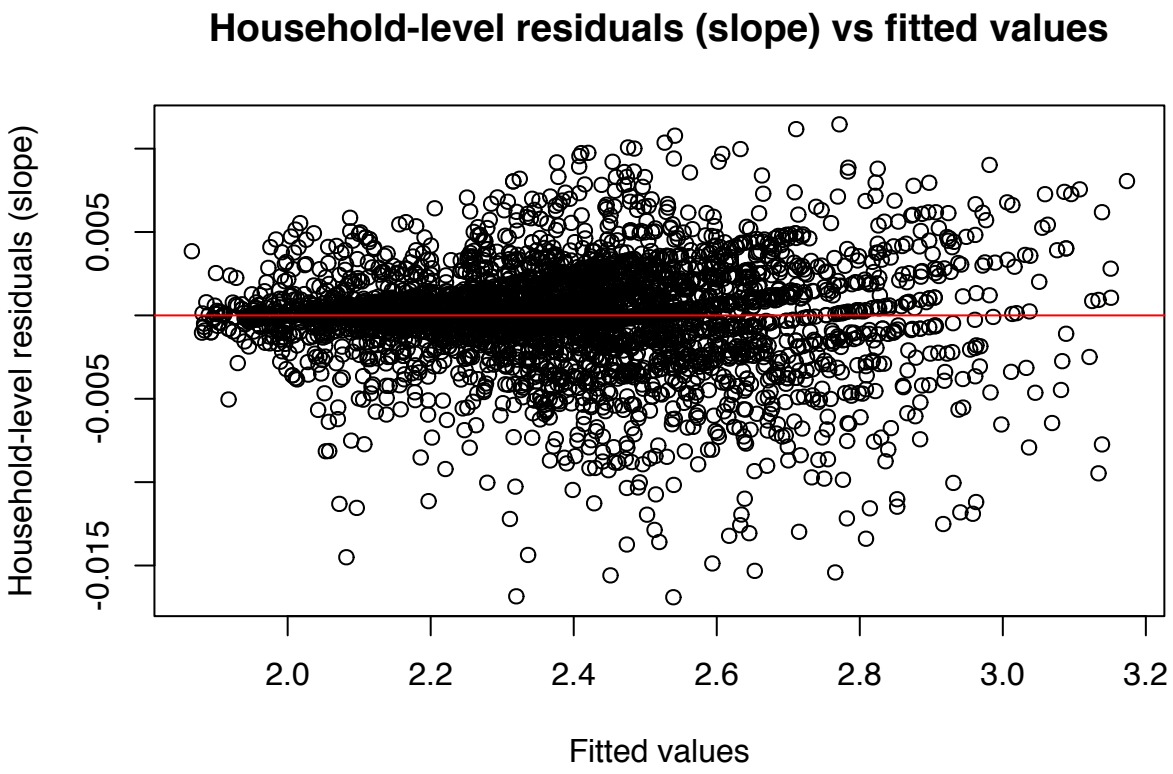
First the individual level:



Then the household-level residuals, for the random intercept:



Then the household-level residuals, for the random slope:



In all three instances, there is cause for concern when looking at the distribution of our residuals.

As with the normality assumption, we find here that the homoscedasticity assumption is stretched, if not outright violated.

Both the q-q plots and the heteroscedasticity plots show that the assumptions of the regression model are not fully met. Whilst they do not seem to be strongly violated, caution should be used when interpreting the model.

Part B

Question 11. Start from the single level null model and add in the household and then GOR levels. Display the model coefficients for the 3 level variance components model in a table.

	Single level null			Two levels (household)			Two levels (region)			Three levels		
Predictors	*Odds Ratios*	*std. Error*	*p*	*Odds Ratios*	*std. Error*	*p*	*Odds Ratios*	*std. Error*	*p*	*Odds Ratios*	*std. Error*	*p*
(Intercept)	0.73	0.03	**<0.001**	0.72	0.03	**<0.001**	0.71	0.07	**<0.001**	0.71	0.08	**<0.001**
Random Effects												
σ^2				3.29			3.29			3.29		
τ_{00}				0.07_{hid}			0.05_{region}			0.06_{hid}		
										0.05_{region}		
ICC				0.02			0.02			0.03		
N				3883_{hid}			12_{region}			3883_{hid}		
										12_{region}		
Observations	4064			4064			4064			4064		
R^2 Tjur	0.000			0.000 / 0.022			0.000 / 0.016			0.000 / 0.034		

Question 12. Is a three level model appropriate? Is a 2 level model appropriate? Is a single level model appropriate? If a 2 level model is appropriate, which two level models (individuals within HH or individuals within GOR)? State your reasons for choosing between a 2 vs 3 level model.

Conducting likelihood ratio tests against the single level null model we find a non-significant improvement in fit for the two-level model with household at level 2 (the test statistic is 0.6337), but a significant improvement in fit for the two-level model with region at level 2 (the test statistic is 41.5984).

Therefore a two-level model would be appropriate with region at level two, but not with household. A three-level model provides a significant improvement in fit when compared to the single level null model (the test statistic is 42.0264) and to the two-level model with household at level 2 (the test statistic is 41.3927), however the improvement in fit is not significant compared to the two-level model with region at level 2 (the test statistic is 0.428).

Both the three-level model and the two-level model with region at level 2 show improvements in fit over the other models, however since there is no statistically significant difference in fit between this model, the most appropriate model is the more parsimonious two-level model with region at level 2.

Question 13. Having decided whether a multilevel or single level model is appropriate, add in the explanatory variables from your final model in Part A. Compare the models using a Table. Are the associations with worry_crime similar to the associations with nhood_mistrust?

	Logistic model		**Linear model**	
Predictors	*Odds Ratios*	*std. Error*	*Estimates*	*std. Error*
(Intercept)	1.17	0.20	2.44	0.06
age	0.99	0.00	-0.00	0.00
sclfsato	0.90	0.02	-0.06	0.01
urban	1.49	0.17	0.34	0.05
school_qual	0.85	0.07	0.07	0.02
no_qual	0.79	0.12	0.01	0.03
rent_local_auth	0.82	0.17	0.43	0.05
rent_private	0.99	0.19	0.17	0.06
ageXurban	1.00	0.00	-0.00	0.00
ageXrent_local_auth	1.00	0.01	-0.01	0.00
ageXrent_private	1.00	0.01	-0.00	0.00
Random Effects				
σ^2	3.29		0.13	
τ_{00}	0.04 $_{region}$		0.39 $_{hid}$	
τ_{11}			0.00 $_{hid.age}$	
ϱ_{01}			-0.71 $_{hid}$	
ICC	0.01		0.66	
N	12 $_{region}$		3883 $_{hid}$	
Observations	4064		4064	
Marginal R^2 / Conditional R^2	0.025 / 0.036		0.117 / 0.703	

There are some similar associations.

For example being in an urban area is associated with higher mistrust in ones neighborhood, and also higher odds of being worried about crime; being satified with life is associated with lower odds of being worried about crime, as well higher levels of neighborhood mistrust.

Variables such as age and the interaction terms for age have very small effects in both models.

However there are also some divergences between models. For example the the education and housing tenure type dummy variables are associated with higher levels of neighborhood mistrust, but lower odds of being worried about crime.

Question 14. Take out any non-significant explanatory variables from the model. Interpret the coefficients in your final model in a table.

	Logistic model		
Predictors	*Odds Ratios*	*std. Error*	*p*
(Intercept)	0.97	0.15	0.850
sclfsato	0.90	0.02	**<0.001**
urban	1.60	0.08	**<0.001**
school_qual	0.84	0.07	**0.012**
no_qual	0.72	0.11	**0.003**
Random Effects			
σ^2	3.29		
$\tau_{00\ region}$	0.04		
ICC	0.01		
$N_{\ region}$	12		
Observations	4064		
Marginal R^2 / Conditional R^2	0.023 / 0.034		

According to this model, having only school-level qualifications, no qualifications, and high levels of satifcation with life is associated with lower odds of being worried about crime. However being in an urban area is associated with higher odds of being worrid about crime ($\approx 8/5$).

Bibliography

Bates, Douglas; Martin Maechler, Ben Bolker, Steve Walker (2015). 'Fitting Linear Mixed-Effects Models Using lme4.' *Journal of Statistical Software*, 67(1), 1-48.

Lüdecke D (2020). 'sjPlot: Data Visualization for Statistics in Social Science'. R package version 2.8.4, <URL: https://CRAN.R-project.org/package=sjPlot>.

R Core Team (2013). 'R: A language and environment for statistical computing.' *R Foundation for Statistical Computing*, Vienna, Austria. Available online: http://www.R-project.org/.

Sarkar, Deepayan (2008) 'Lattice: Multivariate Data Visualization with R'. *Springer*, New York.

Wickham, Hadley (2016). 'ggplot2: Elegant Graphics for Data Analysis'. *Springer*, New York.

Xie, Yihui (2020). *TinyTeX: A lightweight, cross-platform, and easy-to-maintain LATEX distribution based on TEX Live.* Available online: https://yihui.org/tinytex/