Fitting complex models using Health Survey data

Clare Salmond
02 June 2006

Contents

Recommendations 2
The experimental model 3
The problems 3
The sampling issues 4
The technical issues concerning weighting 4
  Single level models 4
  Multi-level models 6
  Post-survey adjustments 6
  Conclusions 7
Selection weight approximations 7
  Single level models 9
  Two-level models 10
  Conclusions 10
Multi-collinearity 10
  A note on collinearity diagnostics 11
  Conclusion 12
Software solutions 12
  Conclusions 14
Fitting general linear models 15
Notes about the supplied weights 15
  The final weight 15
  Use of weights 16
  Use of the replicated weights 16
Appendix 1: SAS 17
  PSU8 17
  Single-level model: SURVEYLOGISTIC 18
  Two-level model: GLIMMIX 18
  Jackknife estimates 18
  SAS procedures: capabilities 19
Appendix 2: MLwiN 20
  MLwiN: capabilities 21

Recommendations

There is good reason not to use the final weight supplied by the Ministry of Health for the purposes of model estimation. Instead, the proposed solution uses 20 dummy variables to approximate the survey selection process and thus make a suitable adjustment to the standard errors of the estimates – under the usual implicit survey assumptions that non-responders are like responders, and that there is no differential non-response. Inference is then from the sample to the target population, as is common. As these implicit assumptions are unlikely to be true, some caution needs to be exercised in interpreting the statistics (as always).

The proposed approach is to include key likely-confounders / exposures-of-interest, plus three more variables indicating survey design, all as dummy [class/nominal] variables in unweighted analyses. These variables (all described in more detail on page 3) are:

confounders/exposures of interest		
Age	use the 4 life-style age groups	3 dummy variables
Sex		1 dummy variable
Ethnicity	this is also a *design variable*	3 dummy variables
design variables		
Stratum		6 dummy variables
PSU8	this is a grouped PSU-size variable	7 dummy variables
Adults in household	Q277, missing categories removed	4 dummy variables

The proposed approach can be used for **linear logistic models**, in single or two-level frameworks, with or without random intercepts, using SAS (proc GLIMMIX), and also for **general linear models** (continuous outcome variable), using, for example, proc GLM. These can also be run in MLwiN but with reservations about (a) the adequacy of the convergence possible, and (b) some small, but non-negligible, differences in the estimated parameter values which may indicate inadequate algorithms. MLwiN users should probably check their proposed final models in a SAS run. In *all* cases the estimated standard errors are likely to be a little smaller than they should be, but this may be of little practical importance. **Single-level linear logistic models** can also be run using SAS proc SURVEYLOGISTIC, and **single-level general linear models** using SAS proc SURVEYREG. In these SAS cases, if the design-variables above are used as well as cluster and strata statements, the standard errors should be close to accurate.

If any other *canonical*[1] **generalised linear models with random intercepts** are desired – such as models with Poisson errors and a log link – they can be done using GLIMMIX or MLwiN. If random slopes as well as random intercepts are desired, only MLwiN can cope (but with reservations as above). However, if *hybrid* **models** are desired – such as using a log link but Binominal errors, to estimate relative risks rather than odds ratios – only GLIMMIX can cope.

For **general linear models**, using either MLwiN or SAS proc SURVEYREG – and, if desired, for **single-level linear logistic models** using SAS proc SURVEYLOGISTIC – a model using the MoH-supplied final weights, which incorporate post-survey adjustments, can be run *without* using the three *design-only* variables, thus saving 17 parameters (but watch out for multi-collinearity, especially if ethnicity is included). Inference then goes from the sample-adjusted-for-non-response to the target-population (with the ususal implicit assumptions). In these cases the standard errors will be too small and a jackknife process (see *Appendix 1*) should be used to empirically estimate appropriate standard errors for the otherwise-'correct' parameter estimates.

---

[1] *Canonical models*: e.g. Normal errors and identity link; Binomial errors and logit link; Poisson errors and log link. *Hybrid models* use a non-canonical-model link for a given error distribution.

The experimental model

Models of smoking and social fragmentation have been investigated in depth. The variables are:

Smoking Q141, binary, recoded 0, 1 (4 missing values), dependent variable

Social fragmentation using sf01_fdec, the full soc frag deciles based on 2001 CAUs (category 1 is the reference category) , main 'exposure' variable

controlling for

(i) *survey-design variables* (to try to avoid weighting)

(a) Stratum_no (selection variable, 1-7; is NOT equivalent to ethnicity)
(b) (i) PSU_no (the clustering variable, if possible, with 1178 values); otherwise simpler ways to codify the psu clusters, such as
(ii) 'psu-size', the number of observations in a PSU (47 categories)*, or
(iii) 'psu8', psu-size grouped into 1-9, 10-19, etc. (8 categories)*
(c) number of adults in the household (1 to 5)

* the SAS code is given in *Appendix 1*, page 17

(ii) *usual confounding variables*

(a) Age – Lifecycle_agegrps has four groups: 15-24, 25-44, 45-64, 65+; and there is also Age (in years) and Agegrp10yr (in 10-year groups)
(b) Sex (M, F)
(c) Ethnicity (using ETH_MPAO) – also a design variable
(d) Nzdep01_quintiles (1-5; least to most deprived)

Specific confounding variables for this analysis are not included here as they are unnecessary for an understanding of the technical issues of interest. Such variables may be exposure to indoor tobacco smoke at work, Q152 (recoded 0, 1; 17 missing values), and Language, Q267, (recoded 0, 1; no missing values).

The problems

A large number of real models were explored

(a) to identify the effect of using a weight statement in various packages

(b) to establish if any package can do a multi-level model in a generalised/general linear model framework with the sampling design of the MoH survey

(c) to identify a work-around in the event that no package can actually do all that we want

(d) to establish the statistical adequacy of the work-around, exploring multi-collinearity issues, practicalities such as convergence issues and numbers of parameters, and model-building issues such as the possible need for interaction terms.

The sampling issues

We want to cope with a sampling design

- that has *clustering* in the selected Primary Sampling Units (labelled PSU though, in fact, they are meshblocks), which means that our *effective* sample is smaller than the *actual* one (people in clusters are more like each other than people from different clusters) – and standard errors should therefore be larger than those for a random sample

- that was *stratified* (PSUs were sampled from a roughly ethnically-stratified sampling frame [on the basis of a variously-defined 'high' proportion of specific ethnic groups]

- in which *households* were selected with different probabilities, including two for which a census was taken (Chatham and Pitt Islands)

- in which *individuals* were selected from households (of varying sizes), one per household

- and for which the only weight information we have is a *final weight* incorporating (a) selection probabilities of PSU, (b) selection probabilities of house-within-PSU, (c) selection probability of one person within the house, and (d) some post-survey adjustments. The purpose of (d) is to allow good estimates of NZ-population parameters and is essential for estimating population totals.

In a framework of *generalised linear models* [linear models generalised to cope with non-Normal errors], we want to consider multi-level models

(a) that have random intercepts for psus (level 2 in MLwiN parlance)
(b) that may have random slopes, too, but I suspect this level of complication may not be possible in either the social fragmentation or CRAINZ analyses (too many parameters)
(c) so that we can contrast area-level effects with individual-level effects.

The framework for a *general linear model* [Normal errors] is considered here as a special case of estimating a generalised linear model.

The technical issues concerning weighting

The major technical problems concerning weighting are:

(a) Can weighting be incorporated?

(b) Is such weighting accurate?

Single level models

Some canonical generalised linear models can be accomplished using SAS (only models with a logit link and Binomial errors in proc SURVEY LOGISTIC), and in MLwiN (Binomial errors or Poisson errors, each with their canonical links – logit and log, respectively – but not hybrid models such as log links and binomial errors).

SAS will cope correctly with a single-level model of stratified, clustered, and *selection probability-weighted* samples via proc SURVEY LOGISTIC, a procedure based on the survey design package SUDAAN. The estimates and their estimated standard errors are both OK.

MLwiN will cope with selection probability-weighted samples (and clustering in two-level models) but the results (in this logistic framework) are dubious. The weights are of particular concern, as shown in the comparable models 1 and 2 in table 1, where we can safely pretend for this purpose of comparison that the weights are 'pure' selection weights at a single level. While both models may have issues of near-multi-collinearity (discussed more-fully later), they ought to produce similar, if inappropriate, estimates. They do not.

Models 3 and 4 introduce control for clustering. Again the models differ substantively. The two SAS models do not differ by much and this is almost-certainly due to the fact that the cluster information per age/sex/ethnic group will be the same information as that contained in the *selection* probabilities. The additional information that is, in fact, in the final weight has apparently not introduced any near-multi-collinearity issues because the standard errors are not unduly inflated in model 3 when compared to model 1.

Table 1: Comparison of single-level logistic models of smoking and social fragmentation in SAS and MLwiN, with weights

					Sex (F)		age 15-24		age 25-44		45-64	
	software	**strata**	**cluster**	**weights**	est	se	est	se	est	se	est	se
1	SAS	no	no	yes	-0.086	0.065	1.200	0.132	1.297	0.112	0.954	0.116
2	MLwiN	no	no	yes	0.062	0.045	1.314	0.090	1.374	0.075	0.946	0.078
3	SAS	no	stmt	yes	-0.086	0.066	1.200	0.134	1.297	0.115	0.954	0.113
4	MLwiN	no	yes*	yes	0.059	0.052	1.124	0.115	1.305	0.100	0.944	0.104

* this has level 2 identified as psu in order to account for clustering in the 1178 psus

Regardless of the above comparisons, from theoretical considerations neither package could cope correctly with these complex samples when the weights comprise both selection probabilities *and post-survey adjustments for non-response*. The reasoning for this is that the post-survey weighting process introduces a level of empirical uncertainty beyond the theoretical sampling errors (so that standard errors should increase a bit more than that due solely to selection weights).

Even SURVEY LOGISTIC and other SURVEY procedures in SAS cannot cope with this added level of uncertainty in their calculation of standard errors, although the estimates themselves should be OK. This is because the standard errors for the sampling issues are based on theory, but the post-sampling adjustments are purely empirical and the theory does not extend to them. Therefore, if we wish to draw inferences directly from the sample (allowing for stratification, clustering, and *selection* probabilities), then the SAS procedures are fine. In the Health Survey analyses, we could do this if we had the (intermediate) selection probabilities, and not just the final weights. *But if we wish to draw inferences from the sample (adjusted for non-response) to the full population (thus using the final weight, the only one we have) then empirical procedures to produce appropriate standard errors are necessary. This is true whichever package or procedure is used.* The (group) jackknife procedures are considered the best to estimate the variances of the (most likely, correct) parameter estimates. This is discussed in more detail in *Appendix 1*.

Multi-level models

In SAS, proc SURVEY LOGISTIC cannot add a random intercept. To run such a two-level model you have to use proc GLIMMIX (which calls proc MIXED repeatedly). GLIMMIX can cope with many flavours of canonical and hybrid generalised linear models, but does not explicitly incorporate clustering effects (effectively reducing the sample size), unlike the specifically designed CLUSTER statement in SURVEY LOGISTIC. Nor, in fact, does MIXED cope with selection probabilities, because there the weight statement is a variance adjustment, not a selection probability. It appears that there is no algebraic connection between the inverse variance weights and the selection probability weights either in terms of the algorithms needed to estimate weighted regressions, or between the weights themselves. Thus we could not adapt proc MIXED to approximately suit our purpose by using clustering and final weights.

In MLwiN, we can easily run a two-level model and add a random intercept. We can add other multi-level refinements as well, if we wish. But the use of weighting is problem, as shown earlier, regardless of whether the weights are appropriate or not. Furthermore, MLwiN is set up to use explicit sampling weights at *each* level, although a rather messy pseudo-decomposition of final weights has been suggested (see the MLwiN Help file). However, like SAS, it does not cope with the non-sampling adjustments incorporated in the final weight.

Post-survey adjustments

Post-survey weighting assumes, implicitly, that the information we have in the sample responders is basically the same as that among the non-responders. [This is also assumed when we draw inferences from the sample responders to our target population without post-stratification adjustments, which has been common practice.]

If the response rate is more-or-less constant among the demographic-characteristic groups that we use to re-create the target population – e.g age/sex/ethnicity, which would probably give 56 sub-groups [7x2x4] – then the selection weights would be equal to the final, adjusted, weights. However, in the Health Survey it is almost certain that the non-response varies by demographic sub-group. *One issue to be addressed in any analytic strategy, therefore, is whether there are any inherent dangers in using demographic sub-sample information to estimate the larger demographic sub-group.* For example, if there is just one respondent in one of the sub-groups, then that information is assumed to represent all the wider group, which is unlikely; and, furthermore, there is no variation postulated in that group at all. In the event that inferences from such a sub-group are deemed unwise then the strategy should address whether the analysis proceeds without that sub-group and the write-up annotated accordingly. But because we do not have access to the selection weights we cannot judge the relative unevenness of response rates across demographic sub-groups, and thus judge whether we have some exceptionally-weighted (selection plus high non-response adjustment) responders. Including such responders in any analysis should increase standard errors accordingly, but will not unless we have adequate jackknife samples from which to estimate standard errors empirically. Including such responders could also lead to a false sense of adequacy.

Robert Templeton at MSD has provided a quick macro to produce jackknife estimates (see page 18); and the MoH final weight replicates could be used. So the process can be automated to a large extent (in SAS), but it does mean that any model has to be run 101 times! Note that this process could only be safely adopted for single-level models using SAS proc SURVEY LOGISTIC or proc SURVEY REG, that is, models where weights are treated appropriately.

It would therefore be sensible to try to create a satisfactory work-around that adequately approximates the sampling issues without the use of any weight statement. The suggestion (from various sources) is that linear terms be added to our models that, as near as possible, mirror the sampling design process. This would be tantamount to approximating the selection weight and thus the package-produced theory-based standard errors would be approximately correct. (But note that inferences would then be made from the sample to the target population with the implied assumptions that the non-responders and responders are alike and that there is no differential non-response.)

Conclusions

(a) Final weighting cannot be incorporated correctly by either package.
(b) If a weight statement is used with final weights, it would always underestimate the standard errors.
(c) If we had the actual selection weights instead of the final weights we could get correct parameter estimates and their estimated standard errors from SAS for single level models.
(d) If a SAS proc SURVEY LOGISTIC weighted single-level model is used (without the design variables as well) we should take the almost-final weighted model and run a jackknife procedure to estimate the standard errors of the parameter estimates. These can be set up via a SAS macro.
(e) For all other models we should build our models using sampling-design variables among the explanatory terms to approximate selection probability information.

Selection weight approximations

In order to identify how well simple linear combinations of available sample information can be used to represent the final weight (given that we don't know the selection weight – bother!), ordinary regressions with final weight as the dependent variable are shown in table 2. All the independent variables are entered as a set of dummy variables. Two approximate ways of adjusting for clustering are used, each variable using dummy variables for its categories. The variables are psu47 and psu8, where each classifies an observation by the size of its particular (geographic) cluster, actual in the case of psu47 and grouped in the case of psu8. The reasoning behind their use is that the size of a cluster may reflect the degree to which its component observations are more alike than a randomly selected group of the same size.

Model 8 includes all the sampling information, as described in the MoH document, as linear terms, except for the *actual* probabilities of selection, since that would require knowledge of denominator populations (for strata, PSUs), as far as can be ascertained. The model explains 79 percent of the overall variation in the final weight. The remaining 21 percent of the variation is due to variations in PSU size (and thus minor variations in selection probabilities of individuals), some other small variations in probabilities of selection due to the over-sampling process, and the adjustments made for non-response. The latter may have been considerable as the overall non-response was 28 percent and, inevitably, it would have varied, at least somewhat, by age/sex/ethnic group.

The various models in table 2 aim to indicate the most practical approximation to the selection weights. Interpretation is complicated by the fact that we do not know the selection weights, only the final (adjusted) ones!

Table 2: Multiple regression models for Finalwgt

	Dependent variables (confounders/design/post-survey adjustments; all as dummies)*								
	design	*design*	*?design*	*post-survey*	*post-survey*	*design*	*design*	*post-survey*	
		confounder		*confounder*	*confounder*			*? confounder*	
Model	**Stratum**	**Ethnicity**	**Interaction**	**Sex**	**Age group**	**Adults**	**PSU**	**NZdep**	**R^2**
1	yes	yes		yes	yes - 4				0.56
2	yes	yes		yes	yes - 4		yes - 47		0.60
3	yes	yes	str*PSU-47	yes	yes - 4		yes - 47		0.62
4	yes	yes	str*eth	yes	yes - 4		yes - 47		0.62
5	yes	yes	PSU-47*eth	yes	yes - 4		yes - 47		0.63
6	yes	yes		yes	yes - 4			yes - 5	0.57
7	yes	yes		yes	yes - 4		yes - 47	yes - 5	0.62
8	yes	yes		yes	yes - 7	yes - 5	yes - 1178	yes - 10	**0.79**
9	yes	yes		yes	yes - 7	yes - 5	yes - 47	yes - 10	0.72
10	yes	yes		yes	yes - 4	yes - 5	yes - 47	yes - 10	0.72
11	yes	yes		yes	yes - 4	yes - 5	yes - 47	yes - 5	0.71
12 ?use	yes	yes		yes	yes - 4	yes - 5	yes - 8	yes - 5	0.70
13 use	**yes**	**yes**		**yes**	**yes - 4**	**yes - 5**	**yes - 8**		**0.68**
14	yes	yes		yes	yes - 7	yes - 5	yes - 47	yes - 5	0.72
15	yes	yes		yes	yes - 7	yes - 5	yes - 47		0.70
16	yes	yes		yes	yes - 4		yes - 1178		0.71
17	yes	yes		yes	yes - 4		yes - 47		0.60
18	yes	yes		yes	yes - 7		yes - 47		0.61
19	yes	yes		yes	yes - 7		yes - 47	yes - 5	0.63
20	yes	yes		yes	yes - 7		yes - 47	yes - 10	0.63
21	yes	yes		yes	yes - 4		yes - 47	yes - 10	0.62

* The numbers attached to each variable indicate the number of categories.

Note: Models 8 and 16 each took over 45 minutes to run! The rest took a few seconds. And these were ordinary regressions – no iterations required. The models are not practical propositions!

If we compare the pairs of models in table 2, we can draw the following conclusions:

(i) changing from NZdep deciles to NZdep quintiles makes little difference to the proportion of variance explained (R^2) (models 10, 11 and 19, 20)

(ii) changing from lifestyle age-groups (4) to 10-year age-groups (7) makes little difference (models 9, 10 and 20, 21)

(iii) changing from (impractical) 1178 psus to 47 psu-sizes [number of responders in the psu]* makes a difference of 0.07 to 0.11 in the proportion of variance explained (models 8, 9 and 16, 17) *defined in *Appendix 1* in the SAS code for PSU8

(iv) changing from 47 psu-size variables to 8 grouped-psu-size variables makes little difference (models 11, 12) (and saves 39 dummy variables)

(v) introducing the household size variable does make a difference: 0.09 in the proportion of variance explained (models 14, 19; 15, 18)

(vi) introducing some interactions makes a small difference of 0.02 to 0.03 (model 2 vs models 3, 4, 5)

(vii) not using a psu design variable makes a difference of 0.04 - 0.15 (models 1, 2 and 1, 16)

These observations suggest that **model 13 (or, possibly, model 12) should be the preferred modelling approach**. They account for a sizeable (and acceptable) proportion of the final weight (which we know includes elements of post-survey adjustment) and does not need vast numbers of dummy variables. [Note that introducing NZDep(5) helps a little, but at the cost of four more dummy variables, so could be left out if there are convergence problems and no other modelling reason to include them.]

The approximation might be improved by the addition of interaction terms, but at the cost of ever-increasing numbers of parameters to estimate, so interaction terms could be introduced into those models where there is enough power. However, I suggest that they be ignored, at least until a near-final model is run. At that point interactions could be introduced but I suspect the added uncertainty involved in introducing many more dummy variables is likely to be undesirable, and multi-collinearity might also become a problem.

Single level models

Table 3 explores the relative merits of using the 46 dummies of PSU-size or the 7 dummies for its eight groups in a generalised model. Note that none of these models includes the number of adults in the household, thus removing possible interpretation problems that might occur from the extra dummy variables needed. These experimental models were all accomplished in SAS.

Table 3: Selected single-level logistic models of smoking and social fragmentation controlling for age, sex, ethnicity and selected design indicators*

					Sex (F)		age 15-24		age 25-44		45-64	
model	**strata**	**cluster**	**adults**	**weight**	est	se	est	se	est	se	est	se
SAS Proc SurveyLogistic												
1	fixed	psu47	fixed	no	0.058	0.045	1.423	0.096	1.476	0.078	1.013	0.080
2	fixed	psu8	fixed	no	0.051	0.045	1.430	0.095	1.483	0.077	1.019	0.080
3	stmt	stmt	no	yes	-0.086	0.066	1.200	0.135	1.297	0.115	0.954	0.114
4	stmt	no	no	yes	-0.086	0.065	1.200	0.132	1.297	0.112	0.954	0.116
5	fixed	stmt	no	no	0.063	0.046	1.315	0.092	1.378	0.078	0.934	0.081
6	fixed	psu8	no	no	0.062	0.044	1.307	0.090	1.379	0.075	0.935	0.078
7	fixed	psu47	no	no	0.070	0.045	1.304	0.090	1.374	0.076	0.932	0.079

* indicators were either a set of dummy variables ('fixed') or a SAS statement ('stmt');
psu<*x*> had *x-1* dummy variables while strata had 7 and adults had 4.

Reassuringly, the standard errors in models 1 and 2 vary by a maximum of 0.001, while the parameter estimates vary a little more, by a maximum of 0.007 – but this is to be expected given the extra 39 dummy variables in model 1.

Thus the suggested strategy is to use 20 survey-design dummy variables to represent the variations in selection probabilities – 6 for strata, 7 for PSU grouped-size, 4 for the number of adults in the household from which the respondent was selected, and 3 for ethnicity (to allow for the complications of over-sampling).

Model 3 would be correct if the final weight was in fact just the overall selection weight. Because it has the added adjustment for non-response, etc, which has been treated only as a further selection weight, the properly-increased standard errors in model 3 are still somewhat too *small* – they do not account for the empirical uncertainty introduced by the non-response adjustments.

There is a niggling concern about the correct adjustment for clustering. Standard errors are increased in the presence of substantial clustering, because the relative 'alikeness' *within* clusters reduces the effective sample size (the square root of which is in the denominator of usual standard error estimators). The effect of the proper treatment of clustering (by a cluster statement, model 3) can be judged by comparing it to model 4. It appears that the added effect of clustering (beyond that introduced, essentially, by the weight) is minimal. The effect of the proper adjustment for clustering can also be judged from the three unweighted models 5, 6 and 7 . The effect is minimal, the maximum discrepancy in standard errors being 0.003. Again, the slightly larger discrepancies between the parameter estimates in the models – 0.011 or less – is most likely fully due to the 7 (or 46) extra dummy terms being estimated in model 6 (or 7).

Two-level models

There is no further issue as far as using linear terms for the design-variables are concerned.

Conclusions

(a) We can use 20 survey-design dummy variables to represent the variations in selection probabilities – 6 for strata, 7 for PSU grouped-size, 4 for the number of adults in the household from which the respondent was selected, and 3 for ethnicity (to allow for the complications of over-sampling).

(b) The additional effect of clustering within PSUs is quite minor when demographic confounders are included in the model. Fixed effect methods of accounting for this clustering (using psu8) are acceptable.

Multi-collinearity

Early explorations suggested the possibility of near multi-collinearity between the ethnicity and stratum dummy variables, even though all ethnicities are represented in each stratum. (The near-collinearity probably occurs with the three much-smaller strata.) This led to an examination of the final weight by ethnicity. Mean final weights are shown below for the 12519 respondents with complete data for the smoking and Social Fragmentation analyses.

group	n	mean	sd	min	max
Asian	1172	134.4	154.3	1.73	750.00
Maori	4120	69.4	132.8	1.53	750.00
Pacific	908	127.5	141.7	1.20	750.00
Other	6329	327.3	209.7	2.62	750.00

Clearly, final weight is related to ethnicity, as you would expect from the sample design. Therefore, to put both in an analysis may introduce some element of multi-collinearity, as would be indicated if there were increases in the standard errors of all, or many, of the estimates. Two models that examine this are shown in table 4.

Table 4: Selected single level survey-logistic models of smoking and social fragmentation in SAS

					Sex (F)		age 15-24		age 25-44		45-64	
model	**ethnicity**	**strata**	**cluster**	**weight**	est	se	est	se	est	se	est	se
1	no	stmt	stmt	yes	-0.073	0.063	1.375	0.127	1.407	0.111	1.005	0.111
2	yes	stmt	stmt	yes	-0.086	0.066	1.200	0.135	1.297	0.115	0.954	0.113

Neither model includes the number-of-adults dummy variables.

Since the standard errors in the two models vary by 0.007 or less, there is no evidence that near multi-collinearity with ethnicity is a problem. But is there any other evidence of near multi-collinearity? Consider the models shown in Table 5.

Table 5: Further selected single level survey-logistic models of smoking and social fragmentation in SAS

				Sex (F)		age 15-24		age 25-44		45-64	
model	**strata**	**cluster**	**weight**	est	se	est	se	est	se	est	se
1	fixed	stmt	no	0.063	0.046	1.315	0.092	1.378	0.078	0.934	0.081
2	fixed	psu8	no	0.062	0.044	1.307	0.090	1.379	0.075	0.935	0.078
3	stmt	stmt	yes	-0.086	0.066	1.200	0.135	1.297	0.115	0.954	0.114
4	no	stmt	yes	-0.086	0.066	1.200	0.134	1.297	0.115	0.954	0.113

No model includes the number-of-adults dummy variables, but all include ethnicity.

The SAS model 3 is a theoretically appropriate model, accounting properly for the clustering and the weighting, as long as you pretend that the final weight is a pure selection weight – and for the purpose of comparison, it does not matter whether the final weight is technically correct or not. Clearly the standard errors in models 3, and 4 (provided to show the lack of impact of strata), are larger than the others in the table. (The negative estimate for sex is of no great concern, because the term is far from significantly different from zero.) The cluster statement does not appear to be the driver for increasing the standard errors since the standard errors are not increased in model 1. The driver for the increased standard errors seems to be the weight, a conclusion that can also be drawn from the models shown in table 3. But could an alternative explanation for the increased standard errors in models 3 and 4 be some near-multi-collinearity?

If we examine the standard errors for the (non-significant) social fragmentation dummy variables (not shown) we find that the range of their standard errors in a SAS model with both strata and cluster statements, as well as the three confounder groups (age/sex/ethnicity), but ignoring weighting, is 0.136 - 0.164. When a weight statement is added, the range becomes 0.249 - 0.277. The standard errors have increased 60 - 80 percent. *Clearly, multi-collinearity IS an issue in the presence of weights, at least in this demonstration model relating smoking to social fragmentation.*

A note on collinearity diagnostics

Beware of the SAS collinearity diagnostics. If dummy variables are present, SAS will indicate multi-collinearity but this may only be collinearity between the *reference groups* (whether or not an intercept is included). The collinearity diagnostics only work for *interval-level* explanatory variables (such as age in years, or BMI), not for nominal ones (such as the 0-1 dummy variables).

To look for multi-collinearity, you should examine the standard errors and check for 'large' increases in standard errors as more variables are added to a model, step-wise fashion. Experience will indicate what, if any, variables may cause a problem. In a large data set like the Health Survey, increases of 50% or more might indicate some near-multi-collinearity that needs to be examined and removed if possible (by removing explanatory variables). Collinearity is more likely with dummy variables representing lots of small groups (so particularly likely if interaction terms are added). The problem is only of importance to inferences among the exposure variable of interest, but is a problem to good-model building if it occurs in *any* variable. Of particular concern might be the stratum dummy variables. The stratum sizes are **64**, 307, 834, **38,** 2695, **68**, and 8323 records respectively, so strata 1, 4, and 6 should be watched, especially if there are lots of parameters in the model.

Conclusion

Multi-collinearity could be an issue in some models, particularly for dummy variables representing small groups. Standard errors in stepwise models should be monitored for sudden increases which would indicate such a problem and suggest either (i) an inappropriate model, and/or (ii) unreliable parameter estimates and/or confidence intervals (or hypothesis tests). Thus models might start with only the exposure of interest, then add the usual age/sex/ethnicity likely-confounder variables, and then add the stratum, psu8, and number of adults dummies. In the event that there appears to be some multi-collinearity, remove any non-significant exposure variables. You could then try removing confounders one at a time to see if their effect is essentially negligible. The survey-design variables need to remain, so the next step, if one is needed, that might help would be to analyse the data in separate subgroups. On the other hand you might just have to accept the uncertainty (high standard errors), in the interests of epidemiological soundness (adding confounders) and statistical soundness (keeping design variables) .

Software solutions

The software

Two packages are readily available to us: SAS and MLwiN.

SAS is well-established and can deal with a variety of general and generalised linear models (including logistic), both single-level and two-levels, through proc GLIMMIX, as well as survey logistic models through proc SURVEY LOGISTIC. GLIMMIX can incorporate random intercepts through a 'random' statement, but it cannot deal with selection weights. SURVEY LOGISTIC can cope with (pure) selection weights, but cannot add a random intercept term.

MLwiN is more recent and the generalised linear model components are fairly new (and probably less-well tested). It copes explicitly with level-2 clustering, but its treatment of weights is problematic, and especially problematic if the selection probabilities are not explicit at *each* level. A complicated method allows you to de-compose a final weight into level-based pseudo-weights but the results are also problematic.

Practical solutions

As described earlier, we can use design-variable information as linear terms (dummies) in our model instead of using a weight: stratum, a grouped-size indicator for the psus (psu8), the number of adults in the selected household, and ethnic group.

We should also include the usual potential confounder variables of age-group, sex, and ethnicity. These will have the added benefit of allowing, in part, for some of the post-survey adjustments, perhaps along with the quintiles of NZDep if that seems warranted in particular models.

We can build our models by including these variables. Our usual interpretations then hold – we draw inferences about the target population from the sample making the usual assumptions that differential non-response bias is negligible, and non-responders are not too dissimilar from responders.

In single level models, SAS proc SURVEY LOGISTIC and MLwiN will give similar results. This is shown for two sets of models in table 6 – models 1 and 2, and 5 and 6 – where the slight differences in estimated standard errors (0.004 or less) are most likely due to the easier convergence criteria used in MLwiN – 10^{-3}, as opposed to the default 10^{-8} in SAS.

Table 6: Comparison of (essentially) single level unweighted models of smoking and social fragmentation in SAS, proc SURVEY LOGISTIC, and MLwiN

					Sex (F)		age 15-24		age 25-44		45-64	
	software	**strata**	**cluster**	**adults**	est	se	est	se	est	se	est	se
1	SAS	fixed	psu8	fixed	0.051	0.045	1.430	0.095	1.483	0.077	1.019	0.080
2	MLwiN	fixed	psu8	fixed	0.051	0.044	1.430	0.094	1.484	0.076	1.020	0.079
3	SAS	fixed	psu8 *	fixed	0.051	0.046	1.430	0.101	1.484	0.080	1.020	0.083
4	MLwiN	fixed	psu8 * †	fixed	0.051	0.044	1.430	0.094	1.484	0.076	1.020	0.079
5	SAS	fixed	stmt	no	0.063	0.046	1.315	0.092	1.378	0.078	0.934	0.081
6	MLwiN	fixed	yes †‡	no	0.063	0.044	1.315	0.089	1.378	0.074	0.934	0.077

* Model 3 has an *additional* 'cluster' (and 'strata') statement. Model 4 has level 2 identified as 'psu' [*not* psu8] *in addition* to psu8. The additions account for any clustering in the 1178 psus not already accounted for by the fixed PSU8 dummies.

† Models 4 and 6 could be classed as two-level models but they have no parameter refinements at level 2, so are essentially comparable to the SAS models.

‡ Model 6 has level 2 identified as 'psu', to compare with the SAS model.

Models 3 and 4 provide reassurance that the use of the fixed design effects is coping quite well with the clustering within the 1178 PSUs. Clustering is of concern, of course, because the usual standard errors need to be adjusted upwards in the presence of any clustering effect. Both models include the software-appropriate ways to account for clustering *as well as* the fixed-effect design effects which are all that we will be able to use in SAS proc GLIMMIX runs. The software differences in models 3 and 4 are slightly larger than in the other pairs in this table, although (as one would hope) the parameter estimates agree, at least to 0.001. The SAS standard error estimates are now up to 0.007 larger than the MLwiN ones, suggesting that the specific cluster statement in SAS is allowing for an element of clustering not fully accounted for by the limited number of PSU8 effects.

This is encouraging, for two reasons. First, it is highly unlikely that there is no clustering effect at all, so we would expect the introduction of a cluster statement to have some noticeable effect

given that only seven dummy variables of PSU8 are approximating the full clustering effect. Second, **the additional amount of clustering is small, and unlikely to make a difference to conclusions drawn from any model which does not, or can not, include a specific cluster statement as well as the fixed design-effects.**

It is, however, of concern that the MLwiN estimates in model 4 are unchanged from model 2. This may be due to differences in philosophy between the single-level approach of Survey Logistic, and the multi-level approach of MLwiN which aims, primarily, to *disentangle residuals* at each level. Which is right? I suspect that SAS is more accurate, because it is increasing the standard errors in a way I would expect, unlike MLwiN. Note that the clustering effect may be larger, or smaller, in models involving other outcome and exposure variables. This could be checked with two quick SAS proc Survey Logistic runs for any final combination of variables in a given analysis (regardless of the actual method of producing the final model, say, a proc Glimmix run). One run could be similar to that shown in Appendix 1; the other would be the same run minus the 'cluster' and 'strata' statements. The observed difference in standard errors between the two outputs for the exposure variables of interest is the likely level of unaccounted-for clustering in the proc Glimmix run (unweighted, of course) with fixed design-variable effects. Hopefully, this minimal extra work could give rise to a one-line reassurance in the Methods section of any resulting paper.

In all other types of generalised linear models, SAS proc GLIMMIX can be used to model random intercepts. As long as no weight statement is used, SAS and MLwiN should give similar – but not near-identical – results, judging by the two models in table 7. The differences are not due to the chosen MLwiN estimation method: the IGLS and RGLS estimates for model 2 differ by at most 0.001 (not shown). Thus, again, part of the difference, at least, is likely to be due to the weaker convergence criteria used in MLwiN. Of note is the fact that the MLwiN process would not converge to even 10^{-4}, at least in well over 100 iterations. There were also some models which would not converge to anything beyond 10^{-2}, which leads to a concern about the differences between the algorithms used in SAS and those used in MLwiN.

Table 7: Comparison of two-level unweighted models of smoking and social fragmentation in SAS, proc GLIMMIX, and MLwiN

					Sex (F)		age 15-24		age 25-44		45-64	
	software	strata	cluster	adults	est	se	est	se	est	se	est	se
1	SAS*	fixed	psu8	no	0.057	0.044	1.281	0.086	1.351	0.070	0.911	0.073
2	MLwiN*	fixed	psu8	no	0.057	0.044	1.302	0.090	1.372	0.075	0.933	0.078

* a random statement for psu was used in GLIMMIX, and a random intercept in MLwiN

Conclusions

(a) Design information can be incorporated as dummy variables in both SAS and MLwiN.

(b) Two-level logistic models with random intercepts can be fitted in SAS proc GLIMMIX and in MLwiN but the MLwiN model should probably be treated as approximate.

(c) More complicated hybrid generalised linear models can only be fitted in SAS.

(d) More complex multi-level canonical* models can only be fitted in MLwiN but, given (b) above, the results should be treated with caution. *defined in the footnote on page 2

Fitting general linear models

The same general conclusions about weights apply as those for the generalised linear models discussed in the foregoing:

(a) Use of the final weights is problematic due to the non-sampling component

(b) Use of a weight statement is almost-certainly problematic in any procedure except SAS proc SURVEY REG.

(c) Dummy variables can be used to approximately mimic the survey design selection probabilities in unweighted SAS proc GLIMMIX and MLwiN models, with some reservations about the accuracy of all the MLwiN results, and the SAS standard errors.

(d) A jackknife procedure using the MoH replicate final weights could be used to appropriately estimate standard errors in single-level weighted SAS proc SURVEY REG models that do not contain the purely design-mimicking dummy variables.

Notes about the supplied weights

The final weight

The Ministry document 'Methodology' indicates that the selection of strata and PSUs within them aimed to 'ensure equal probability of selection for every dwelling within the target population'. Selection was then one person per dwelling, so depended on household size (adults) – information in variable Q277. While each stratum indicated a high proportion of a certain ethnicity, selection within a stratum was not restricted to one ethnicity. However, some ethnic groups were over-sampled, so ethnicity is a key design variable (as well as one of interest). Probability of PSU selection varied by strata, so stratum is an important design component.

This suggests that a combination of stratum, PSU, ethnicity, and household size *largely* identified the *sampling strategy* variables. In turn, this suggests that the final weight is composed of stratum, PSU, ethnicity, and household size information, plus extra information on non-sampling errors such as non-response/non-contact, since the final weight was determined to re-produce the population structure of NZ in 2001.

The extra adjustment to produce the final weights does not involve sampling design, so the 'extra' element of error in the weighted parameter estimates is not governed by sampling theory. If we can approximate the sampling (selection) weights by some means other than using a weight, the standard errors of our estimates will allow for the sampling errors introduced by the survey design. To allow also for non-sampling errors (bias), we would have to use an empirical method (such as a group jackknife) which could be rather time consuming for the type of models we want to fit.

Use of weights

In epidemiological-type surveys we normally assume that relationships between variables are similar among the responders and the non-responders as long as key variables are distributed *similarly* in these two groups. Further, we do not usually worry about differential non-response. If we analyse data within the age/sex/whatever groups that might have differential non-response there is, in fact, no problem. If differential non-response is considered at all, the argument is that there are much larger 'errors' in our data than the small adjustments needed to allow for differential non-response. But this argument is inadequate when the purpose at hand is to estimate population totals, which may have fiscal or other important implications. In the powerful modern computing environment, another argument says we should use these more-correct estimation methods whenever feasible.

The key issue for surveys where the main purpose is to investigate *relationships* is whether there is an appreciable *differential* non-response in key groups, because this could affect the relative importance of those groups in determining the relationships (e.g. calculating a correlation coefficient). One way to approach this is to analyse the survey separately within these groups, but this would be cumbersome and we would lose large amounts of statistical power. A preferable approach would be to include the group information in the model. The linear terms may not cope fully with the differential response, although interaction terms might improve matters. The problem with introducing interaction terms, of course, is the plethora of extra variables needed, especially if nominal variables are involved. However, we could build models with linear terms, and check the final model for any change caused by introducing interaction terms as long as we have enough power to do so. The best approach, however, is to use the survey weights and a jackknife method to estimate standard errors. *This pre-supposes that a procedure is available that correctly utilises the weights in the first place.* As we have seen, this is only applicable in special circumstances for the sorts of models we want to explore.

We have shown that we can not use weights safely in either MLwiN (see table 1), or in SAS proc GLIMMIX (see *Multi-level models*, page 6). The standard errors are wrong in both cases, and the parameter estimates are suspect in MLwiN. We can use weights in SAS proc SURVEY LOGISTIC, or proc SURVEY REG, for single-level models, as long as there is no major near-multi-collinearity. Thus, only in the SAS 'SURVEY' procedures could we use the jackknife procedure and the MoH-supplied replicated weights.

Use of the replicated weights

The replicated weights [Finalwgt_rep1, etc.] were created by the delete-a-group form of the empirical jackknife procedure.

Using the replicated weights, 100 replicated analyses could be used to estimate the standard errors of the parameter estimates, where the final standard deviation of an estimate is the standard deviation of the 100 estimates of the parameter. In this way we attempt to control adequately for the various forms of sampling and non-sampling error in our sample data.

Suggestion: It might be instructive to run one such model just to see how much difference it actually makes over and above the suggested use of 20 dummy variables to describe the main features of the survey design. I suspect it makes little difference (because the inferences from sample-to-population using the usual theory-based standard errors, and sample-adjusted-for-non-response-to-population using empirical [jackknife] standard errors, assume the same things, namely non-responders are like responders, and there is no differential non-response). If so, it would provide both reassurance and an opportunity for a smug statement in a methods section!

Appendix 1: SAS

NOTE:
Please contact June if you have any queries or difficulties in using/adapting these codes.

Sample code for creating the variable **PSU8** *and running the procedures* **SurveyLogistic** *and* **Glimmix**.

```
*----CREATE PSU8----;

libname s 'E:\Test'; *----change as required----;
x      'cd E:\Test';

*----create cluster size [psu_total] from 1178 unique PSU numbers----;
*----NOTE: a numeric version of the character variable psu_no (psu) had
     been made in an earlier data step using psu = inputn(PSU_no,8.);

proc  sort   data=s.nzhs2003_curf_nz_v2c;         by psu_no;
proc  means  data=s.nzhs2003_curf_nz_v2c noprint; by psu_no;
      output out=b n=psu_total;    *----psu_total has 47 categories----;
      var    psu;
run;


*----create new file and add psu8----;

data  s.newcurf;
merge s.nzhs2003_curf_nz_v2c b; by psu_no;
psu8  = int(psu_total/10)+1;
run;


*----check results----;

proc freq data=s.newcurf;
     tables psu_no*psu_total*psu8 / list;
run;
```

[Sample lines from the cross-tabulation list are shown below.]

PSU_no	psu_total	psu8	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	12	2	12	0.10	12	0.10
2	9	1	9	0.07	21	0.17
3	11	2	11	0.09	32	0.26
4	9	1	9	0.07	41	0.33
5	22	3	22	0.18	63	0.50
6	27	3	27	0.22	90	0.72
7	13	2	13	0.10	103	0.82
8	18	2	18	0.14	121	0.97
9	2	1	2	0.02	123	0.98
10	8	1	8	0.06	131	1.05
.....						
1171	7	1	7	0.06	12450	99.37
1172	12	2	12	0.10	12462	99.47
1173	8	1	8	0.06	12470	99.53
1174	12	2	12	0.10	12482	99.62
1175	12	2	12	0.10	12494	99.72
1176	12	2	12	0.10	12506	99.82
1177	10	2	10	0.08	12516	99.90
1178	13	2	13	0.10	12529	100.00

```
*----EXAMPLES, both set up for convergence to 0.00001, although SAS seems
     to have little difficulty in converging to the default 1E-8----;

*----Single-level model: SURVEYLOGISTIC----;

proc printto new file='TestRuns.out'; run; *---file for selected output---;
ods select none;

ods select type3 parameterestimates;
title1 'Single-level logistic model using proc SurveyLogistic';
title2 'Fixed effects for ethnicity, stratum, psu8, and adults';
proc  surveylogistic data=s.newcurf;
      class   sf01_fdec(ref='1') Lifecycle_agegrps Sex Eth_MPAO stratum_no
              psu8 hhadults / param = ref;
   strata  stratum_no;
   cluster psu;
      model   smoke (event='1') = sf01_fdec Lifecycle_agegrps Sex Eth_MPAO
              stratum_no psu8 hhadults / GCONV=1E-5;
run;
proc printto; run;


*----Two-level model: GLIMMIX----;

title1 'Two-level logistic model using proc Glimmix';
title2 'Fixed effects for ethnicity, stratum, psu8, and adults';
title3 'Random intercept at level-2 (psu)';

proc   glimmix data=s.minicurf2 pconv=1E5;
class  sf01_fdec Lifecycle_agegrps Sex Ethnic stratum_no psu8 hhadults;
model  smoke (event='1') = sf01_fdec Lifecycle_agegrps Sex Ethnic
       stratum_no psu8 hhadults /dist=binary solution;
random int/subject=psu;   *----random int[ercept] at level 2(psu)----;
run;
```

Sample macro for creating **jackknife estimates**.

The following SAS macro has been adapted by June Atkinson from one written by Robert Templeton. June intends to extend the macro to deal with the Survey Logistic procedure, when required.

```
*----Repeatedly run the survey procedure for each jackknife and put
     the results together in one data set----;


%MACRO JACKS (NUM_JACKS=100,DATASET=jacks);

*----NUM_JACKS is the number of jackknife weights you want to use.
     Default=100. DATASET is the data set the analyses are to be
     done on. Default=jacks----;

%DO I=1 %TO &NUM_JACKS;

ODS  OUTPUT STATISTICS=MEAN&I (RENAME=(MEAN=MEAN&I));

PROC SURVEYMEANS DATA=&dataset;
     STRATA      STRATUM_NO;
     CLUSTER     PSU;
     WEIGHT      FINALWGT_REP&I ;
     VAR         <.....>;
RUN;
%END;
```

[macro continued overleaf]

```
DATA ALLMEANS;
  MERGE %DO I=1 %TO &num_jacks ; MEAN&I  %END; ;
  STD_ERR=SQRT((&num_jacks /(&num_jacks-1))* CSS(OF MEAN1-MEAN&num_jacks));
  OVERALL_MEAN=MEAN(OF MEAN1-MEAN&num_jacks);
  LOWER=OVERALL_MEAN- 1.96 *STD_ERR;
  UPPER=OVERALL_MEAN+ 1.96 *STD_ERR;
RUN;

PROC PRINT DATA=ALLMEANS;
  VAR OVERALL_MEAN STD_ERR LOWER UPPER;
RUN;
```

SAS procedures: capabilities

Proc SURVEYLOGISTIC

(i) Logistic models (Binomial errors and a logit link) in a survey setting. Logistic models are a member of the class of generalised linear models. The dependent (outcome) variable is usually binary, such as smoker (yes/no). Independent variables (exposures, confounders, exposure-modifiers, etc.) can be nominal (as one or more dummy yes/no variables), ordinal (ditto, which thus ignores the ordering), or interval (such as BMI or age).
(ii) Allows for clustering via the 'cluster' statement, and stratification (not so crucial) via the 'strata' statement.
(iii) Single-level models only.
(iv) The weight statement allows for *selection* probabilities (but is inadequate for weights that are also adjusted for differential non-response, such as the MoH final weight).
(v) Can be used with Jackknife replicate-weights for selection probabilities adjusted for differential non-response.
(vi) Can also provide suitable models using the design-variable-fixed-effects, no-weights approach.

Proc GLIMMIX

(i) Single-level and simple two-level generalised linear mixed (random and fixed effects) models which can use a variety of links and error distributions. Identity link and Normal errors produces linear multiple regression models. Logit link and Binomial errors produces logistic models. The sample should be simple random, that is, not have a complex survey design. The form of the dependent variable should be consistent with the link and the error-distribution. Thus both general linear models and linear logistic models can be fitted, as well as several others. Permitted types of variables are described in Proc Surveylogistic (i) and Proc Surveyreg (i).
(ii) Can not allow for any non-simple-random survey characteristics (clustering, strata).
(iii) Two-level models can have random intercepts by using the 'random' statement.
(iv) The weight statement assumes the weights to be proportional to the reciprocals of the error variances, or similar, so is not suitable for selection-probability weights.
(v) Since the weight statement is not geared to selection probabilities, in whole or in part, empirical Jackknife estimates of standard errors can not be utilised.
(vi) Can provide suitable models using the design-variable-fixed-effects, no-weights approach.

Proc GLM

(i) Linear multiple-regression models for simple random samples (See Surveyreg (i) for descriptions of variables).
(ii) Can not allow for any non-simple-random survey characteristics (clustering, strata).
(iii) Single-level models only.
(iv) The weight statement assumes the weights to be proportional to the reciprocals of the error variances, or similar, so is not suitable for selection-probability weights.
(v) Since the weight statement is not geared to selection probabilities, in whole or in part, empirical Jackknife estimates of standard errors can not be utilised.
(vi) Can provide models using the design-variable-fixed-effects, no-weights approach.

Proc SURVEYREG

(i) General linear multiple-regression models (Normal errors) in a survey setting. The dependent (outcome) variable is continuous [interval] (e.g. BMI) but not necessarily Normal or even completely symmetric. The residual term, however, is Normal, so any non-Normality in the dependent variable should be accounted for by the explanatory variables. Hence the exploration of the distribution of residuals in good analytic practice.
(ii) Allows for clustering via the 'cluster' statement, and stratification (not so crucial) via the 'strata' statement.
(iii) Single-level models only.
(iv) The weight statement allows for *selection* probabilities (but is inadequate for weights that re also adjusted for differential non-response).
(v) Can be used with Jackknife replicate-weights for selection probabilities adjusted for differential non-response
(vi) Can also provide models using the design-variable-fixed-effects, no-weights approach.

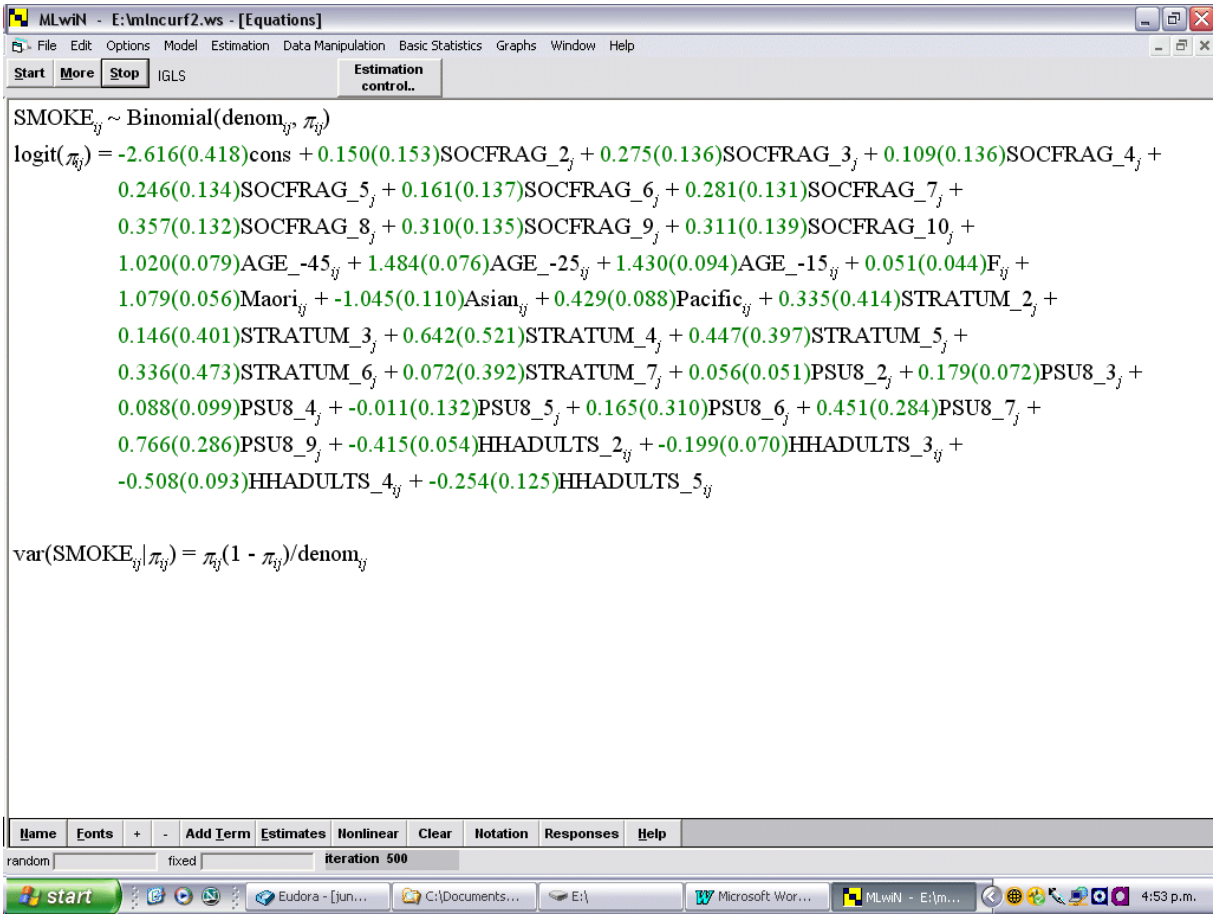
Appendix 2: MLwiN

A screen capture of an MLwiN run is shown on the next page. It has two levels: psu is level-2 (1178 values); and a generated ID of 1 to 12529 is level-1. The model does NOT have a random intercept (see model 4, table 6) thus allowing for clustering at the PSU-level in a single-level model framework.

Estimation control was set to 0.001 (note that the default is 0.01). Occasionally, a model did not converge.

Note: the category variable 'PSU_9$_j$' is not a mistake; it is automatically produced – clearly the MLwiN code does not allow a variable name ending in a number – PSU8 – to have a category member ending in the same number.

The simple, essentially single-level, model on the next page has level-2 defined (subscript j) but does not have a random intercept at level 2.



MLwiN: capabilities

The following description uses the same framework as the descriptions of the capabilities of the various SAS procedures in *Appendix 1*.

(i) General linear multiple-regression models (Normal errors) as well as canonical generalised linear models (e.g. Binomial errors with a logit link; Poisson errors with a log link).

(ii) Allows for clustering via the level-2 identification in the dependent variable definition but, apparently, not adequately.

(iii) Single-level models, two-level models, and more, with allowances for random intercepts and random coefficients. Most flexible modelling across two or more levels.

(iv) The weight adjustments do not appear to be adequate. Furthermore, if used, component weights are required *at each level*, so any overall weight would have to be empirically de-composed into pseudo-weights at each level. The method of decomposition (see MLwiN help file) should be considered as developmental.

(v) Since the weight adjustments are not adequate, empirical Jackknife estimates of standard errors could not be considered reliable.

(vi) Can provide *probably approximate* models using the design-variable-fixed-effects, no-weights approach.