Chapter 13
Untangling Context and Composition

Abstract This chapter contains a tutorial that helps to untangle contextual and compositional effects. We start from a typical, empty table and then proceed to fill this table. The example data set concerns patterns of incidence of cardiovascular disease in small areas in Scotland. The outcome or dependent variable is whether or not a survey respondent had self-reported doctor-diagnosed cardiovascular disease. The first step in the analysis is to estimate a null model. We then estimate the fixed effects of two individual-level variables, social class and smoking status, one by one. The final model looks at the fixed effects of all three variables. With these steps the empty table can be filled and we can interpret the results in terms of context and composition.

In this chapter, we describe the analysis of these data using MLwiN.

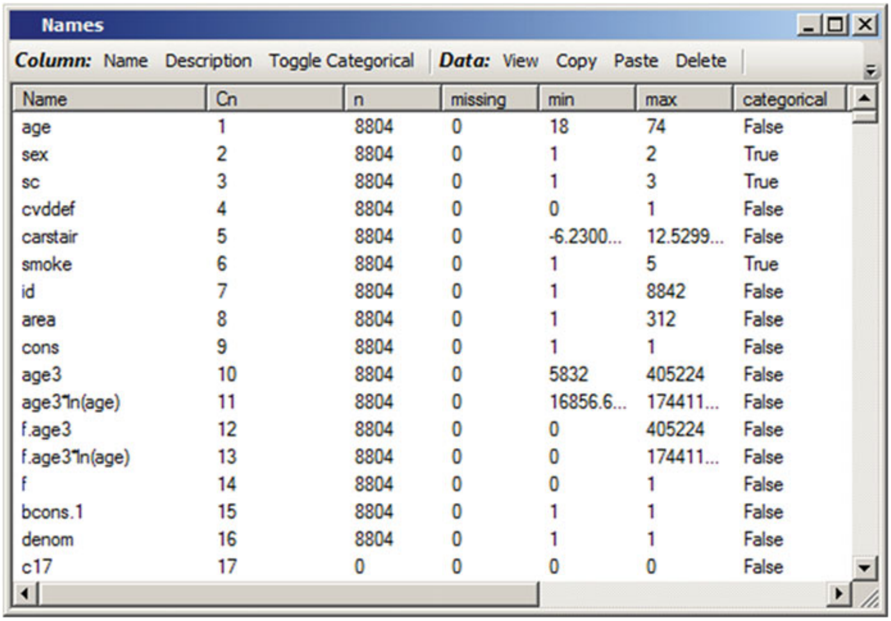
Keywords Tutorial · Multilevel analysis · Compositional effect · Contextual effects · Cardiovascular disease

As we pointed out in Chap. 7, there is frequent debate in the literature over the relative contributions of composition and context in the statistical explanation of individual-level outcomes, such as self-reported health and the incidence and prevalence of disease or mortality. This tutorial provides an application of the insights from Chap. 7. In this tutorial we will be looking at the patterning of the prevalence of cardiovascular diseases in Scotland. In particular, we consider whether the prevalence of disease is related to an individual social determinant (occupational social class), an individual biological determinant (current smoking status) or an area-based social determinant. As an area-based social determinant we used area deprivation measured by the Carstairs score, a Census-based variable derived from the social class of the heads of households, male unemployment, lack of car ownership and overcrowding (Carstairs 1995; Carstairs and Morris 1990). As with the previous two chapters, the software used in this chapter is MLwiN. Further details on multilevel modelling and MLwiN are available from the Centre for Multilevel Modelling http://www.bristol.ac.uk/cmm/. The materials have been written for MLwiN v3.01. The teaching version of the software is available from https://www.bristol.ac.uk/cmm/software/mlwin/download/.

A. H. Leyland, P. P. Groenewegen, *Multilevel Modelling for Public Health and Health Services Research*, https://doi.org/10.1007/978-3-030-34801-4_13

The Data

The data are contained in the worksheet ‘CVD-data.wsz’ and are taken from the 1998 Scottish Health Survey, and the analysis is related to a published paper (Leyland 2005). The data refer to 8804 respondents aged between 18 and 64. The outcome considered is a self-report of a doctor-diagnosed cardiovascular disease (CVD) condition (angina, diabetes, hypertension, acute myocardial infarction, etc.). This is a binary response, whether (1) or not (0) respondents have CVD condition.


Names
Column: Name Description Toggle Categorical
Data: View Copy Paste Delete

Name	Cn	n	missing	min	max	categorical
age	1	8804	0	18	74	False
sex	2	8804	0	1	2	True
sc	3	8804	0	1	3	True
cvddef	4	8804	0	0	1	False
carstair	5	8804	0	-6.2300...	12.5299...	False
smoke	6	8804	0	1	5	True
id	7	8804	0	1	8842	False
area	8	8804	0	1	312	False
cons	9	8804	0	1	1	False
age3	10	8804	0	5832	405224	False
age3*ln(age)	11	8804	0	16856.6...	174411...	False
f.age3	12	8804	0	0	405224	False
f.age3*ln(age)	13	8804	0	0	174411...	False
f	14	8804	0	0	1	False
bcons.1	15	8804	0	1	1	False
denom	16	8804	0	1	1	False
c17	17	0	0	0	0	False

The independent variables at individual level on which we focus in the tutorial are social class and smoking status. Occupational social class is used in three categories: high social class (1 and 2: professional and managerial), intermediate (3: skilled workers), and low (4 and 5 and missing: semiskilled and unskilled manual workers and those for whom social class was missing). Smoking has been categorised as never smoked, light smokers (<10 cigarettes per day), moderate (10–19) and heavy (20+) smokers as well as former smokers. Age and sex are used as control variables in all analyses. At the area level the Carstairs index is used as a continuous variable.

The survey was cluster-sampled, with respondents clustered within 312 small areas (postcode sector, average population about 5500).

Structure of the Analysis

As a first exploratory step in the analysis, examine the mean Carstairs score by social class and current smoking, and also smoking patterns by social class, to see the dependency between the variables.

After that, we are going to examine a series of models with a view to determining the relationship between the prevalence of CVD diseases and individual social class, current smoking and area deprivation. We will conduct these analyses with a table in mind, filling in the table as we progress (see Table 13.1).

Estimating the Null Model

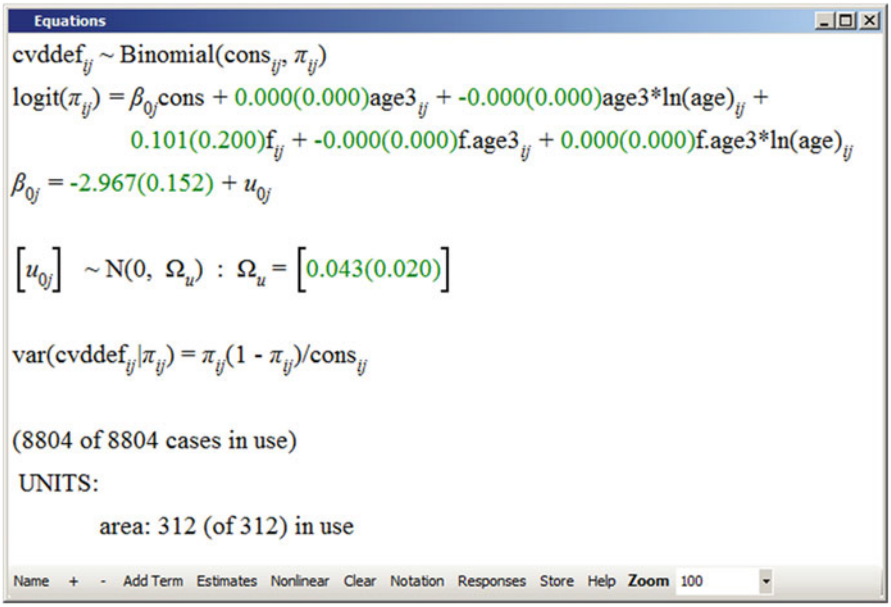
The first model to fit is a null model. We will adjust all of the models we fit for age and sex, but we are not going to report the estimates associated with these factors; these are 'nuisance variables' and we are going to control for differences between areas in their age and sex composition.

We then set up a two-level model with the response variable CVDDEF and with levels defined by AREA and ID. This is a binomial response with a logit link function and with the denominator given by the constant CONS. We will add CONS to the fixed part of the model and allow for random intercepts across areas by letting the coefficient of CONS vary at random at level 2 (i.e. across areas). It is important that we have a well-fitting model at individual level, otherwise unmeasured individual effects might appear as contextual effects. We have used fractional polynomials in age (Royston et al. 1999) together with interactions with sex to find a parsimonious model that adequately controls for age and sex; these are already included in the model that can be found in the **Equations** window. We can start off by fitting this model using the first order MQL approximation but then move on to the second order PQL approximation. This is then the null model on which we base subsequent analyses.

Table 13.1 Outline of a table to report the analysis to untangle context and composition

		Null		Social class		Smoking		Deprivation		Social class + deprivation		Smoking + deprivation	
Variable		OR	CI	OR	CI	OR	CI	OR	CI	OR	CI	OR	CI
	Fixed part												
Social class													
	1&2			Baseline						Baseline			
	3												
	4&5												
	p												
Smoking													
	<10												
	10 < 20												
	20+												
	Ex-smoker												
	Never smoked					Baseline						Baseline	
	p												
Area deprivation													
	p												
	Random part												
Area variance													
Individual variance		[a]		[a]		[a]		[a]		[a]		[a]	
ICC													
R^2													

[a] Individual variance for multilevel logistic regression models approximated by $\pi^2/3$ (Snijders and Bosker 2012)


Equations
Name + - Add Term Estimates Nonlinear Clear Notation Responses Store Help Zoom 100

$$\text{cvddef}_{ij} \sim \text{Binomial}(\text{cons}_{ij}, \pi_{ij})$$

$$\text{logit}(\pi_{ij}) = \beta_{0j}\text{cons} + 0.000(0.000)\text{age3}_{ij} + -0.000(0.000)\text{age3*ln(age)}_{ij} + 0.101(0.200)\text{f}_{ij} + -0.000(0.000)\text{f.age3}_{ij} + 0.000(0.000)\text{f.age3*ln(age)}_{ij}$$

$$\beta_{0j} = -2.967(0.152) + u_{0j}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim \text{N}(0,\ \Omega_u)\ :\ \Omega_u = \begin{bmatrix} 0.043(0.020) \end{bmatrix}$$

$$\text{var}(\text{cvddef}_{ij}|\pi_{ij}) = \pi_{ij}(1 - \pi_{ij})/\text{cons}_{ij}$$

(8804 of 8804 cases in use)

UNITS:

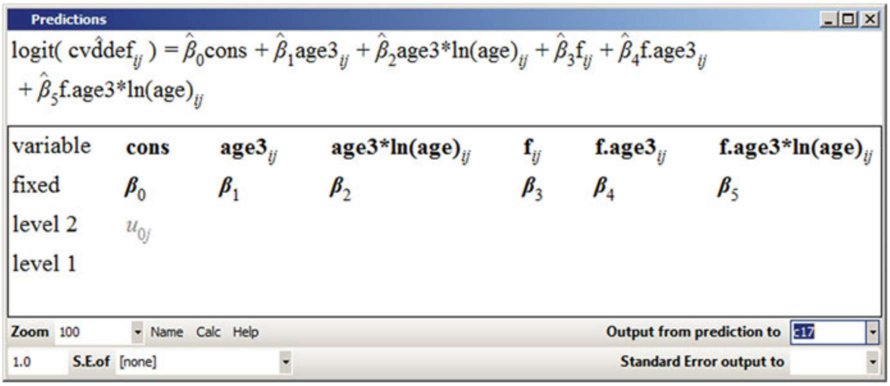
area: 312 (of 312) in use

We can estimate the ICC from this model using the approximation that the individual-level variance is given by $\pi^2/3$ (= 3.290). So a level 2 variance of 0.043 gives an ICC of 0.013; just over 1% of the variation in the prevalence of CVD diseases is attributable to differences between areas.

A useful diagnostic measure is the *R*-squared which indicates how much of the total variation has been explained by the fixed part of the model. For multilevel logistic regression, we approximate the explained variation by the variance of the linear predictor (that is, the variance of the fixed part of the model which is on a log odds scale) and get the total variance by adding the variance of the linear predictor to the variance at the higher levels plus our estimate of the variance at the individual level. In other words,

$$R^2 = \text{VLP}/\left(\text{VLP} + \sigma^2_{u0} + \pi^2/3\right)$$

where VLP is the variance of the linear predictor. We can calculate the linear predictor using the Predictions window and including all variables in the fixed part (but not the random part).



$$\text{logit}(\,\hat{\text{cvddef}}_{ij}\,) = \hat{\beta}_0\text{cons} + \hat{\beta}_1\text{age3}_{ij} + \hat{\beta}_2\text{age3*ln(age)}_{ij} + \hat{\beta}_3\text{f}_{ij} + \hat{\beta}_4\text{f.age3}_{ij} + \hat{\beta}_5\text{f.age3*ln(age)}_{ij}$$

variable	cons	age3$_{ij}$	age3*ln(age)$_{ij}$	f$_{ij}$	f.age3$_{ij}$	f.age3*ln(age)$_{ij}$
fixed	β_0	β_1	β_2	β_3	β_4	β_5
level 2	u_{0j}					
level 1						

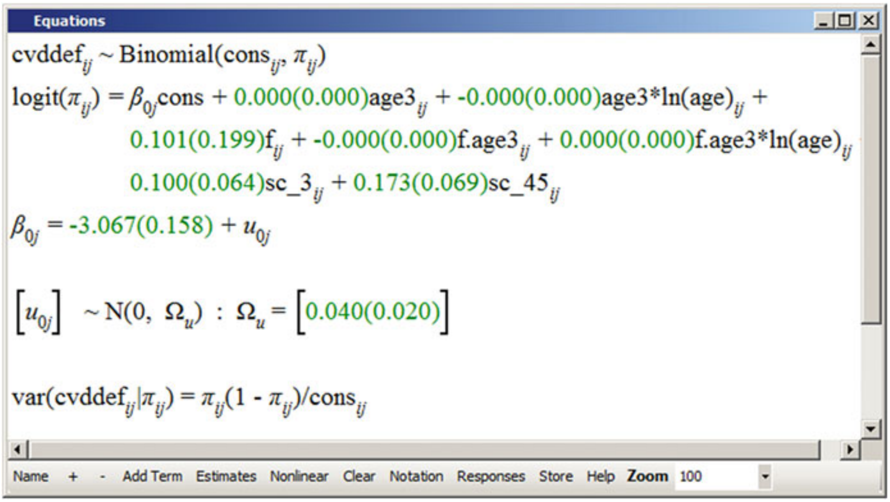
Zoom 100 · Output from prediction to c17
1.0 · S.E.of [none] · Standard Error output to

We can use the **Averages and correlations** window to estimate the standard deviation of this prediction as 0.921. The variance is the square of the standard deviation; this gives VLP = 0.848 and so *R*-squared = 20.3%.

The values of the ICC, VLP and *R*-squared can be obtained for any two-level multilevel logistic regression model by running the macro ‘modeldiag.txt’. (To run the macro make sure that the output window of the **Command interface** is open, then open the macro using the **File** menu and click **Execute**.)

Fixed Effects

The first model that we want to fit is the model containing individual social class (variable SC). There are three categories of social class; we will fit two dummy variables keeping social class 1 and 2 as the reference category.



$$\text{cvddef}_{ij} \sim \text{Binomial}(\text{cons}_{ij}, \pi_{ij})$$

$$\text{logit}(\pi_{ij}) = \beta_{0j}\text{cons} + 0.000(0.000)\text{age3}_{ij} + -0.000(0.000)\text{age3*ln(age)}_{ij} + 0.101(0.199)\text{f}_{ij} + -0.000(0.000)\text{f.age3}_{ij} + 0.000(0.000)\text{f.age3*ln(age)}_{ij}\ 0.100(0.064)\text{sc\_3}_{ij} + 0.173(0.069)\text{sc\_45}_{ij}$$

$$\beta_{0j} = -3.067(0.158) + u_{0j}$$

$$\left[u_{0j}\right] \sim \text{N}(0,\ \Omega_u)\ :\ \Omega_u = \left[0.040(0.020)\right]$$

$$\text{var}(\text{cvddef}_{ij}|\pi_{ij}) = \pi_{ij}(1 - \pi_{ij})/\text{cons}_{ij}$$

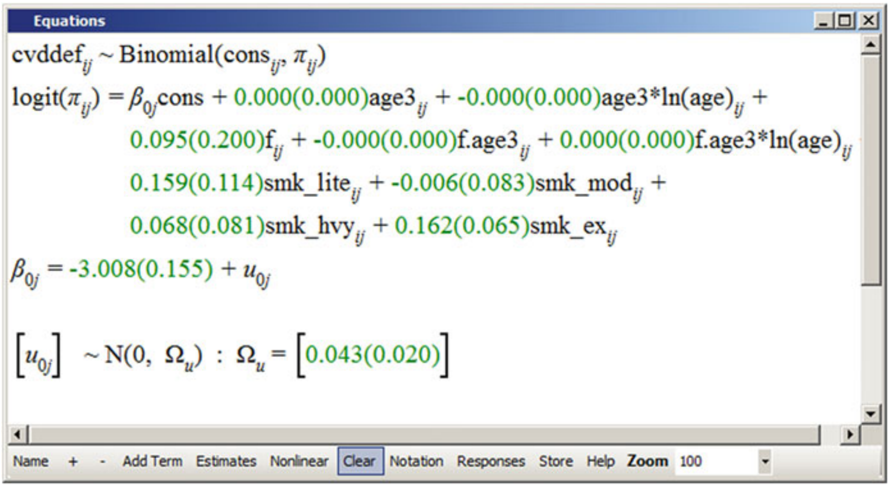
Zoom 100

The parameter estimate for social class 3 is a log odds ratio; we can convert this to an odds ratio by exponentiating: $\exp\{0.100\} = 1.105$, so the odds of CVD diseases are 10.5% higher in social class 3 than in social classes 1 and 2. Similarly we can obtain 95% confidence intervals as $\exp\{0.100 \pm 1.96 \times 0.064\} = (0.975, 1.253)$. Since the 95% confidence interval for this odds ratio includes 1, it suggests that the odds ratio for social class 3 is not significantly different from that for social classes 1 and 2.

Odds ratios and 95% confidence intervals can be obtained for all parameter estimates from any logistic regression model by running the macro ‘or.txt’.

Although the odds ratio for social class 3 is not significantly different from that for social classes 1 and 2, that for social classes 4 and 5 is significant (the 95% confidence intervals do not include 1). Since we would expect the social class effect to increase across social class categories—CVD prevalence is likely to be higher in social class 3 than in social classes 1 and 2, and higher still among social classes 4 and 5 than in social class 3—we test for a linear trend in the social class variable. We do this by removing the categorical social class variable from the model, fitting social class using a continuous variable created for this purpose (i.e. with values 1, 2 and 3) and testing for the significance of this single variable. This can be done using the **Intervals and tests** window from the **Model** menu.

We can now continue by fitting models containing just smoking and just deprivation (again including age and sex as these were contained in the null model). (Click on a variable in the **Equations** window and choose **Delete term** to remove it from the current model.)



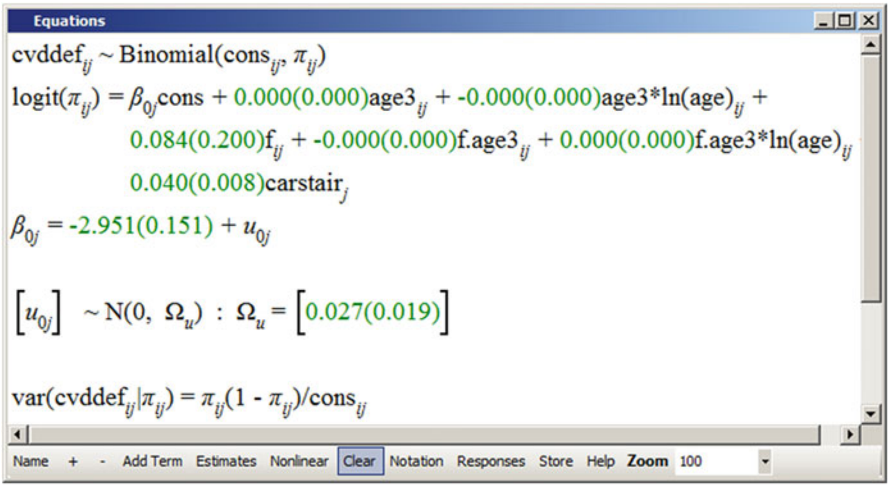
$$\text{cvddef}_{ij} \sim \text{Binomial}(\text{cons}_{ij}, \pi_{ij})$$

$$\begin{aligned}\text{logit}(\pi_{ij}) = \beta_{0j}\text{cons} &+ 0.000(0.000)\text{age3}_{ij} + -0.000(0.000)\text{age3*ln(age)}_{ij} + \\ &0.095(0.200)\text{f}_{ij} + -0.000(0.000)\text{f.age3}_{ij} + 0.000(0.000)\text{f.age3*ln(age)}_{ij} \\ &0.159(0.114)\text{smk\_lite}_{ij} + -0.006(0.083)\text{smk\_mod}_{ij} + \\ &0.068(0.081)\text{smk\_hvy}_{ij} + 0.162(0.065)\text{smk\_ex}_{ij}\end{aligned}$$

$$\beta_{0j} = -3.008(0.155) + u_{0j}$$

$$\left[u_{0j}\right] \sim \text{N}(0,\ \Omega_u)\ :\ \Omega_u = \left[0.043(0.020)\right]$$

Compared to the reference group of never smokers, the prevalence of CVD diseases is no higher in any of the smoking categories but is significantly higher among the ex-smokers. As a prevalence study this may reflect an increased likelihood of giving up smoking once a respondent has been told by a doctor that they have a cardiovascular disease. The categories of smoking are not ordered and so testing the significance of this variable involves testing the significance of differences between categories rather than a test for trend.



$$\text{cvddef}_{ij} \sim \text{Binomial}(\text{cons}_{ij}, \pi_{ij})$$

$$\text{logit}(\pi_{ij}) = \beta_{0j}\text{cons} + 0.000(0.000)\text{age3}_{ij} + -0.000(0.000)\text{age3*ln(age)}_{ij} +$$

$$0.084(0.200)\text{f}_{ij} + -0.000(0.000)\text{f.age3}_{ij} + 0.000(0.000)\text{f.age3*ln(age)}_{ij}$$

$$0.040(0.008)\text{carstair}_{j}$$

$$\beta_{0j} = -2.951(0.151) + u_{0j}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim \text{N}(0,\ \Omega_u)\ :\ \Omega_u = \begin{bmatrix} 0.027(0.019) \end{bmatrix}$$

$$\text{var}(\text{cvddef}_{ij}|\pi_{ij}) = \pi_{ij}(1 - \pi_{ij})/\text{cons}_{ij}$$

Area deprivation is coded with positive values indicating areas of higher deprivation and negative values indicating areas of lower deprivation. The effect of deprivation is clearly significant; we can consider whether the effects of social class and smoking are significant after controlling for area deprivation. At the same time we will see whether the effect of area deprivation remains significant once individual factors are taken into account. The significant effect of individual social class is attenuated and becomes non-significant when area deprivation is taken into account whilst area deprivation remains significantly related to the prevalence of CVD diseases. The effect of individual smoking status remains insignificant following adjustment for area deprivation.

Basically, with these models we can complete Table 13.1 such that it becomes Table 13.2. This presents a neat summary of the fixed and random parts of the models that we have fitted. The strong influence of the context can be seen through the persistent significance of the area deprivation score even after adjustment for individual factors.

Table 13.2 Estimates from model

		Null		Social class		Smoking		Deprivation		Social class + deprivation		Smoking + deprivation	
Variable		OR	CI	OR	CI	OR	CI	OR	CI	OR	CI	OR	CI
	Fixed part												
Social class													
	1&2			Baseline						Baseline			
	3			1.11	(0.97, 1.25)					1.07	(0.94, 1.21)		
	4&5			1.19	(1.04, 1.36)					1.12	(0.97, 1.28)		
	p			0.012						0.123			
Smoking													
	<10					1.17	(0.94, 1.47)					1.15	(0.92, 1.43)
	10 < 20					0.99	(0.84, 1.17)					0.95	(0.80, 1.12)
	20+					1.07	(0.91, 1.26)					1.01	(0.86, 1.19)
	Ex-smoker					1.18	(1.03, 1.34)					1.18	(1.03, 1.34)
	Never smoked					Baseline						Baseline	
	p					0.093						0.053	
Area deprivation													
								1.04	(1.02, 1.06)	1.04	(1.02, 1.06)	1.04	(1.03, 1.06)
	p							0.000		0.000		0.000	
	Random part												
Area variance		0.043		0.040		0.043		0.027		0.026		0.026	
Individual variance		[a]		[a]		[a]		[a]		[a]		[a]	
ICC		0.013		0.012		0.013		0.008		0.008		0.008	
R^2		0.203		0.203		0.204		0.207		0.207		0.208	

[a] Individual variance for multilevel logistic regression models approximated by $\pi^2/3$ (Snijders and Bosker 2012)

Additional Models

There are a variety of other models that we may wish to fit. One of the reasons for the closer relationship between the Carstairs score and the prevalence of CVD diseases may be because the Carstairs score is a continuous variable—indicating a broad range of deprivation—whilst our measure of occupational social class is categorical with just three categories. To satisfy our curiosity that this is not just a measurement issue, we can categorise the deprivation measure into three approximately equal groups and fit some of these models again.

As we discussed in Chap. 3, contextual variables may be direct observations made on areas detailing, for example, the provision of services. They may be derived from alternative data sources (as in this case: the Carstairs score is based on Census variables). Another possibility is to create contextual variables through the aggregation of individual variables collected in the study. Think about creating a contextual variable describing the social class of the neighbourhood. A simple example would be the proportion of the survey respondents in each area who were in social classes 4 and 5; an alternative might be the difference between the proportion in social classes 4 and 5 and the proportion in social classes 1 and 2. Such variables can be created using the **Multilevel data manipulations** window found under the **Data manipulation** menu. These variables permit further examination of the relative importance of composition versus context, given that both descriptors are derived from the same source, but also illustrate how an important contextual descriptor can be created within the data set in the absence of an externally validated measure such as the Carstairs score.

The aggregation of an individual variable to an area level can change its interpretation. We can construct an area-based smoking score to illustrate this. If an individual is given a score of 3 for a heavy smoker, 2 for a moderate smoker, 1 for a light smoker and 0 for an ex-smoker or a non-smoker, then the average of this score at an area level provides information about current smoking behaviour in an area in terms both of smoking prevalence and dose. The relationship of such a variable to the prevalence of CVD diseases is different to the relationship between individual smoking behaviour and CVD disease prevalence; the area smoking score—just like the area social class score—acts as a marker of area deprivation.

References

Carstairs V (1995) Deprivation indices: their interpretation and use in relation to health. J Epidemiol Community Health 49(Suppl 2):S3–S8

Carstairs V, Morris R (1990) Deprivation and health in Scotland. Health Bull 48:162–175

Leyland AH (2005) Socioeconomic gradients in the prevalence of cardiovascular disease in Scotland: the roles of composition and context. J Epidemiol Community Health 59:799–803

Royston P, Ambler G, Sauerbrei W (1999) The use of fractional polynomials to model continuous risk variables in epidemiology. Int J Epidemiol 28:964–974

Snijders TAB, Bosker RJ (2012) Multilevel analysis: an introduction to basic and advanced multilevel modeling. Sage, Los Angeles