Using Quasi-variance to Communicate Sociological Results from Statistical Models

Vernon Gayle and Paul S. Lambert

Sociology 2007; 41; 1191
DOI: 10.1177/0038038507084830

Sociology
Copyright © 2007
BSA Publications Ltd®
Volume 41(6): 1191–1208
DOI: 10.1177/0038038507084830
SAGE Publications
Los Angeles, London,
New Delhi and Singapore

Using Quasi-variance to Communicate Sociological Results from Statistical Models

Vernon Gayle
University of Stirling

Paul S. Lambert
University of Stirling

ABSTRACT

This article introduces a sociological audience to 'quasi-variances' as a solution to the 'reference category problem'. The reference category problem is associated with the interpretation of the effects of categorical explanatory variables within statistical models, and is especially relevant to sociological applications, where categorical explanatory variables are very common. This article presents a selection of examples (using multiple and logistic regression) to illustrate and comment on quasi-variance calculations for sociological models. In addition, the article is augmented with online materials provided by the authors, which aim to help social researchers practise and apply this technique using the popular data analysis software packages SPSS and Stata. The authors conclude that quasi-variance methods offer an attractive and practicable solution to the reference category problem that can, and should, be routinely operationalized by sociological researchers.

KEY WORDS

categorical variables / quasi-variance / reference category / regression models / statistical models

Introduction

Many readers will be familiar with statistical modelling approaches to the analysis of survey data. In particular, sociologists have tended to use regression models in order to explore the effects of multiple explanatory variables on an outcome of interest.

Sociologists are becoming increasingly familiar with statistical modelling techniques. This is partly due to advances in statistical software packages (e.g. SPSS, 2005; Stata, 2005), rapid increases in the power of desk top computers, and the increased accessibility of survey datasets.[1] In addition, sociology postgraduate students are now routinely expected to be trained in common statistical modelling techniques.[2]

Many explanatory variables in social science research are categorical, by which we mean they are measured according to membership of one of a number of discrete categories. Almost all standard statistical models can readily incorporate categorical explanatory variables, and a section explaining standard methods for achieving this has become a regular feature of texts which introduce statistical models to social scientists (e.g. Fielding and Gilbert, 2006: 292–6; for an extended introduction, see Hardy, 1993).

This article is aimed at sociologists who may be engaged in using statistical modelling techniques that include categorical explanatory variables, and at those who read published work that employs such models. In three recent papers British statistician David Firth has advanced a method to assist in the presentation and interpretation of statistical models with categorical explanatory variables, which is termed 'quasi-variance' (Firth, 2000, 2003; Firth and de Menezes, 2004). At the current time, to the best of our knowledge, these papers are not widely known within the British sociological research community and we suspect that this is partly due to their mathematical nature.

This article describes Firth's idea of 'quasi-variance' through simplified sociological examples. We provide a number of examples to illustrate the flexibility of the quasi-variance approach, and focus upon the circumstances when it is most relevant to sociological research. In addition, we have mounted public access data, and a number of example files which use the popular software packages SPSS and Stata, to help sociological researchers practise and apply this technique.

The Reference Category Problem

Statistical models offer an attractive way for sociological researchers to summarize patterns from social survey datasets (Dale and Davies, 1994; Goldthorpe, 2007). They offer techniques to summarize the joint relative effects of several different variables in a research study. This is achieved by estimating statistical values ('parameters' or 'coefficient estimates') that indicate the magnitude and direction of the effect of each explanatory variable. In recent decades, the expansion of statistical methods and data resources in survey research has widened the range of social processes which may be informatively studied through statistical models.[3] Nevertheless, the appropriate sociological interpretation of the parameter estimates from statistical models is by no means trivial (Berk, 2004). Although there are numerous accessible guides to the mathematical interpretation of parameter estimates in social science examples (e.g.

Allison, 1999; Menard, 2001), the important point is that the communication of results from statistical models hinges upon which aspects of the modelling process the analyst chooses, rightly or wrongly, to emphasize (Berk, 2004).

In standard statistical models the effects of a categorical explanatory variable are assessed by comparison to one category (or level) that is set as a benchmark against which all other categories are compared. The benchmark category is usually referred to as the 'reference' or 'base' category. The reference category effect is fixed to zero in the model estimation procedure,[4] and other category effects are interpreted as the additional impact of not being in the reference category. Standard statistical software undertakes formal comparisons of whether or not each category effect differs from the reference category effect. These comparisons generate the well known 'significance values' of parameter (coefficient) estimates.

The reference category problem is easily stated. Whilst it is straightforward to compare any one category with the reference (or base) category, it is more difficult to formally compare two other categories (or levels) of the explanatory variable with each other when neither is the base category. A primary data analyst can calculate formal contrasts between different levels of the same categorical variable. However, the information necessary to undertake these calculations is not usually reported in the outputs of statistical models. Therefore secondary analysts, such as those reading published results, cannot make such comparisons themselves.[5] As we describe later, Firth's papers (2000, 2003) illustrate how 'quasi-variance' statistics can be reported along with standard outputs from statistical models in order to enable such calculations.

Examples

We illustrate the deployment of quasi-variance calculations through a series of survey data analysis examples. The examples draw upon analyses of the UK Census Sample of Anonymised Records (SARs, see ONS, 2005) and an extract from the General Household Survey (GHS) of 2002. We have chosen these datasets because they can be freely downloaded.[6] To accompany these examples we have developed a number of Stata and SPSS syntax files to help readers reproduce these illustrative analyses, and an Excel calculator to assist in statistical calculations. These files, which also include materials referring to further example applications of quasi-variance calculations, can be downloaded from our website (www.longitudinal.stir.ac.uk/qv/) along with a more comprehensive article that elaborates on the issues raised here.

Example 1 and an Introduction to Quasi-variance

The first example (model 1, shown in the first four columns of Table 1) is a logistic regression model using the SARs data. The outcome variable is a binary measure which records whether the person was in good health over the last 12

Table 1 Logistic regression prediction that self-rated health is 'good' (Parameter estimates for model 1, featuring conventional regression results, and quasi-variance statistics)

	1 Beta	2 Standard Error	3 Prob.	4 95% Confidence Intervals		5 Quasi-Variance
No Higher qualifications	–	–	–	–	–	–
Higher Qualifications	0.65	0.0056	<.001	0.64	0.66	–
Males	–	–	–	–	–	–
Females	–0.20	0.0041	<.001	–0.21	–0.20	–
North East England	–	–	–	–	–	0.0000755
North West England	0.09	0.0102	<.001	0.07	0.11	0.0000294
Yorkshire & Humberside	0.12	0.0107	<.001	0.10	0.14	0.0000400
East Midlands	0.15	0.0111	<.001	0.13	0.17	0.0000477
West Midlands	0.13	0.0106	<.001	0.11	0.15	0.0000380
East of England	0.32	0.0107	<.001	0.29	0.34	0.0000394
South East England	0.36	0.0101	<.001	0.34	0.38	0.0000272
South West England	0.26	0.0109	<.001	0.24	0.28	0.0000426
Inner London	0.17	0.0122	<.001	0.15	0.20	0.0000743
Outer London	0.27	0.0111	<.001	0.25	0.29	0.0000480
Constant	0.48	0.0090	<.001	0.46	0.50	–

n = 1,099,214
Log likelihood = –689228.17 (Pseudo-R^2 = 0.015).
Source: UK Census 2001, 3% individual level SARs for England, unweighted.

months (0 = no; 1 = yes). There are three explanatory variables in the model: one for Government Office Region, one for gender and one for education.

We focus our attention on Government Office Region as this provides a simple and clear example of a multiple category explanatory variable with a large number of categories (i.e. 10). In a conventional analysis one region will be set in the model as the reference (or base) category. In this example it is the North East. The parameter estimates (or coefficients) for the other regions are comparisons with the North East. The output reports an estimate for the North West (0.09) that is significantly different to the North East region ($p < 0.001$). Yorkshire and Humberside region is also significantly different to the North East with an estimate of 0.12 ($p < 0.001$).

It is plausible that a reader may wish to make other comparisons between Government Office Regions. For example, a researcher reading the results may wish to establish whether or not the effects of living in the North West and Yorkshire and Humberside are significantly different to each other. From the usual reported outputs (i.e. parameter estimates and their standard errors) it is not possible for the reader to satisfactorily make this comparison.

A comparison is often attempted by using confidence intervals for the parameter estimates. The simple calculation ($\beta \pm (1.96 * \text{standard error})$) can be used to construct a 95% confidence interval around the parameter estimate (β).

In the current example we could, for instance, compare the 95% confidence interval for the North West (0.07 to 0.11) with the equivalent interval for Yorkshire and Humberside (0.10 to 0.14). This is presented graphically through the left-hand markers in Figure 1 (the circles). Since these confidence intervals overlap we might be beguiled into concluding that the two regions are not significantly different to each other. However, this conclusion represents a common misinterpretation of regression estimates for categorical explanatory variables. These confidence intervals are not estimates of the difference between the North West and Yorkshire and Humberside, but instead they indicate the difference between each category and the reference category (i.e. the North East). Critically, there is no confidence interval for the reference category because it is forced to equal zero. A useful analogy is to consider that the confidence intervals for other categories are 'artificially' wider because of this constraint, meaning that these confidence intervals could overlap even when the difference between the categories is significant.

Comparing Categories – Conventional Calculations

Continuing with example 1, in a conventional statistical model we denote the beta estimates for the North East, the North West and Yorkshire and Humberside as β_1, β_2 and β_3 respectively. It is possible to formally test the difference between the North West region and Yorkshire and Humberside by

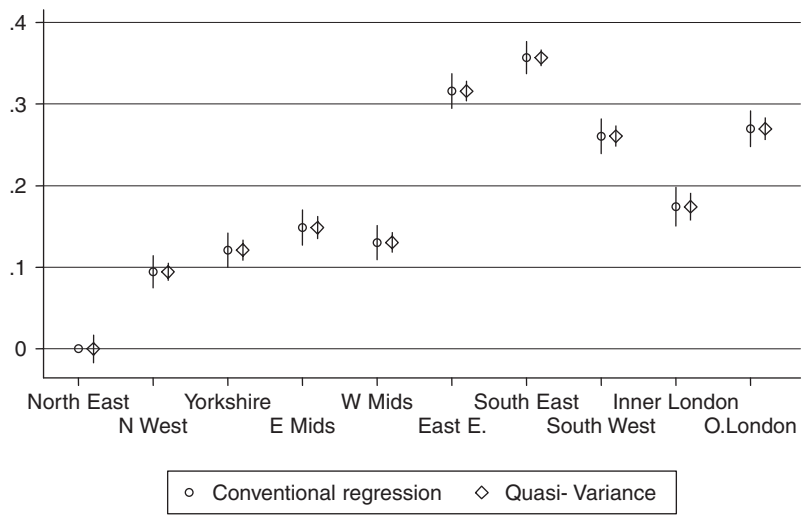

Figure 1 Predictions of Good Health, by Government Office Region Confidence Intervals of regression coefficients, by estimation method

Source: UK Census 2001 SARS for England n = 1,099,294

Model 1: Logistic regression predicting 'Good Health'. Other controls for education and gender

evaluating a *t*-statistic for the unstandardized parameter estimates given in equation 1 (for a detailed discussion see Hardy and Reynolds 2004).[7]

$$t = \frac{\hat{\beta}_2 - \hat{\beta}_3}{\text{s.e. } (\hat{\beta}_2 - \hat{\beta}_3)} \quad (1)$$

It is simple enough to compute the difference between the two beta estimates for the North West and Yorkshire and Humberside (0.09 – 0.12= –0.03, see Table 1). However calculating the standard error of this difference is not as straightforward. The standard error of the difference is conventionally calculated from the following formula:

$$\text{s.e. difference} = \sqrt{\text{var}(\hat{\beta}_2) + \text{var}(\hat{\beta}_3) - 2(\text{cov}(\hat{\beta}_2 - \hat{\beta}_3))} \quad (2)$$

The standard error of the difference between $\hat{\beta}_2$–$\hat{\beta}_3$ therefore requires information on the 'covariance' between the two parameters. This is generated during the estimation of the statistical model, and is conventionally stored in a table known as the 'variance-covariance matrix of the parameter estimates'. Table 2 gives this matrix for example 1. The variance of $\hat{\beta}_2$ can be found in row 1, column 1 of Table 2; the variance of $\hat{\beta}_3$ in row 2, column 2; the covariance between the two parameter estimates can be found in row 2, column 1.

This variance-covariance matrix is not routinely displayed by software in final outputs. It is available in many standard data analysis packages such as Stata, though it cannot be easily displayed for all models in SPSS.[8] With the appropriate covariances, we can make a calculation of the standard error of the difference between the estimate for the North West and Yorkshire and Humber Government Office Regions. For this example:

$$0.0083 = \sqrt{0.00010483 + 0.00011543 - 2(0.00007543)}$$

This calculation then allows us to derive the *t*-statistic:

$$t = -0.03 / 0.0083 = -3.2$$

Using conventional statistical criteria, if the *t* value is greater than ±1.96, we can reject the null hypothesis and conclude that the estimate for the North West is significantly different to Yorkshire and Humberside ($p < 0.05$). For consistency with other standard forms of statistical testing, this calculation should be taken a step further to generate a Wald chi-square statistic (equal to t^2), which is then evaluated at 1 degree of freedom:

$$\text{Wald } \chi^2 = (-0.03 / 0.0083)^2 = 10.22;\ p = 0.0014$$

Table 2 Variance Covariance Matrix of Parameter Estimates for the Government Office Region variable in Model 1

Row	Column	1 North West	2 Yorkshire & Humberside	3 East Midlands	4 West Midlands	5 East England	6 South East	7 South West	8 Inner London	9 Outer London
1	North West	.00010483								
2	Yorkshire & Humberside	.00007543	.00011543							
3	East Midlands	.00007543	.00007543	.00012312						
4	West Midlands	.00007543	.00007543	.00007543	.00011337					
5	East England	.00007544	.00007543	.00007543	.00007543	.0001148				
6	South East	.00007545	.00007544	.00007544	.00007544	.00007545	.00010268			
7	South West	.00007544	.00007543	.00007544	.00007543	.00007544	.00007546	.00011802		
8	Inner London	.00007552	.00007548	.0000755	.00007547	.00007554	.00007572	.00007558	.00015002	
9	Outer London	.00007547	.00007545	.00007546	.00007545	.00007548	.00007555	.00007549	.00007598	.00012356

The value of the Wald χ^2 is significant and we can formally conclude that these two regions are different with regard to self-rated good health.

Recall that this is a different conclusion than would have been reached through the 'eyeballing' of confidence intervals in Figure 1. We reiterate that the erroneous conclusion that might be drawn from Figure 1 arises due to the reference category problem. It occurs because the confidence interval estimates for the North West and Yorkshire and Humberside are comparisons with the North East (i.e. the reference category), which is necessarily set to zero. It is important to appreciate that accurate tests of the contrasts between different levels of a categorical variable are seldom undertaken and reported in sociological outputs.[9]

The key point, however, is that it is ordinarily only the primary analyst who has the opportunity to make formal comparisons between categories. The conventionally reported outputs from statistical models do not include the variance-covariance matrix of the parameter estimates so do not allow the secondary analyst to perform such tests. It is nevertheless prohibitive to expect analysts to routinely publish such matrices, which can be very large in size.[10] Firth's (2003) recommendation, that analysts routinely display 'quasi-variance' statistics for all multiple category explanatory variables, offers a neat and practical solution to this *impasse*.

In essence, Firth's method (2000, 2003) uses an approximation in order to allow for an easier calculation of the test statistic for the difference between two categories.[11] A single approximation statistic, known as the quasi-variance, may be calculated for each category of a categorical explanatory variable (including the reference category). The important outcome is that this statistic may be used to generate a more simplified equation for approximating the standard error of the difference between two beta estimates as used in equation (1). The new calculation for equation (2) becomes:

$$\text{s.e. difference} \approx \sqrt{quasi\ \text{var}(\hat{\beta}_2) + quasi\ \text{var}(\hat{\beta}_3)} \qquad (3)$$

By replacing the expression (2) with (3), as long as the quasi-variance statistic for each beta has been reported, a secondary analyst, for example the reader of a journal article, can readily calculate a *t*-statistic using the conventional formula (1).

The procedure for generating quasi-variances is illustrated in the coming examples, and is repeated in several illustrations on our webpages.[12] Firth provides an online calculator which we use in this illustration:[13]

www2.warwick.ac.uk/fac/sci/statistics/staff/academic/firth/software/qvcalc/web/

To use the online calculator, the analyst must supply two relevant pieces of information on their model estimates. The first is the number of levels of the categorical explanatory variable (in our example this is the 10 Government Office Regions). The second is information from the variance-covariance

matrix of the parameter estimates.[14] In our experience, the precise format of the necessary data from the variance-covariance matrix has confused some colleagues. To help avoid confusion, on our website we illustrate a number of practical examples of extracting the necessary information from the variance-covariance matrix of parameter estimates in applications which use both SPSS and Stata.

The web-based calculator produces a quasi-variance for each level of the categorical explanatory variable.[15] For Example 1, the outputs from the quasi-variance estimates are reported in column 5 of Table 1. With these values, a formal test of the difference between the parameter estimate for the North West and Yorkshire and Humberside can easily be calculated, since the standard error of the difference between the estimates (3) is taken as:

$$\sqrt{0.00010483 + 0.00011543 - 2(0.00007543)} =$$

$$\sqrt{quasi\ \mathrm{var}(\hat{\beta}_2) + quasi\ \mathrm{var}(\hat{\beta}_3)} \quad \sqrt{0.0000294 + 0.0000400} = 0.0083$$

This allows the subsequent calculation of the t and Wald statistics, and the evaluation of the significance of the difference between categories:

$t = (0.09–0.12) / 0.0083 = –3.2$ and

$Wald\ \chi^2 = (–.03 /0.0083)^2 = 10.22;\ p = 0.0014$

The results reported from Firth's quasi-variance approach are identical to the results calculated using the conventional approach based on the variances and covariances of the parameter estimates. This computation may at first seem daunting, so to aid researchers in performing necessary calculations we have constructed an Excel calculator to undertake this estimation online.[16] Using Firth's approximation we would draw the correct conclusion that these two Government Office Regions are different with regard to self-rated good health.[17]

In practice, we have found that a graphical example has helped researchers to better comprehend this issue. In the right-hand markers of Figure 1 (diamond shaped), we have plotted the parameter estimates of model 1 and constructed confidence intervals from quasi-variances (Firth suggests the term 'comparison intervals' for these measures). In Figure 1 we can see that using quasi-variances, the comparison intervals for the North West and Yorkshire and Humberside no longer overlap.

Example 2

In this example we fit a multiple regression model to data from the 2002 General Household Survey (GHS) (see ONS, 2006). The outcome variable is self-reported

Table 3 Multiple regression prediction of age of leaving education (Parameter estimates for model 2)

	Beta	Standard Error	Prob.	Quasi-variance
Age in years – 40	–0.05	0.00	<.001	–
Social class:				
Advantaged (n = 1679)	–	–	–	0.0038
Lower-supervisory (n = 279)	–1.80	0.16	<.001	0.0228
Semi-Routine (n = 524)	–1.93	0.13	<.001	0.0121
Routine (n = 397)	–2.33	0.14	<.001	0.0160
Constant	18.40	0.06	<.001	–

n = 2,879
Log likelihood = –6746.6 (R^2 = 0.203).
Source: 2002 UK General Household Survey, all adults aged 16+, unweighted.

data on the age at which the individual left full-time education (min = 10; max = 50; mean = 17.35; s.e. = 0.05). The model includes explanatory variables for age and social class. We have used the NS-SEC occupational classification based upon current or last occupation (Rose and Pevalin, 2003), but for illustrative purposes we have collapsed it into four categories representing 'advantaged' occupations, lower supervisory occupations, semi-routine occupations and routine occupations. Table 3 gives a conventional presentation of the results from this model,[18] with the additional presentation of quasi-variance statistics.

The reference category for the social class variable in model 2 is 'advantaged' occupations. One interesting question would be whether the age at which those in the 'lower supervisory' classification left education is significantly different to those from 'routine occupations'. Because quasi-variances have been reported, this test can be readily conducted on the basis of the Table 3 outputs. Using the same notations as Example 1:

$$t = \frac{\hat{\beta}_2 - \hat{\beta}_4}{\sqrt{quasi\ \mathrm{var}(\hat{\beta}_2) + quasi\ \mathrm{var}(\hat{\beta}_4)}} = \frac{-1.80 + 2.33}{\sqrt{0.0288 + 0.016}}$$

We can conclude that there is a significant difference in the age at which those from lower supervisory occupations and those from routine occupations leave education. This calculation can easily be undertaken by using the Excel calculator available from our website.

Example 3

Interaction effects in statistical models are often substantively important, but in our experience the effects of interactions are often difficult to communicate

with readers. This is especially acute when dealing with higher order interactions (which by their nature involve many explanatory variables). In this example we tackle the issue of reporting interaction effects and demonstrate that Firth's method is sufficiently flexible to handle them.

Table 4 shows the model outputs from two simple logistic regression models in which the outcome variable is self-reported data on whether or not the respondent reports that they used to smoke regularly, but no longer do so (1 = ex-smoker). We use just two explanatory variables, gender (0 = male, 1 = female), and a definition of age groups designed to highlight the differences between the youngest and oldest sample members (0 = aged 60–69 years; 1 = aged 20–59 years; 2 = aged 16–19 years). The interpretation of the results of models 3.1 and 3.2 is perhaps best aided by a short description of the main model findings. Men are generally more likely to be ex-smokers than women, and older people are generally more likely to be ex-smokers than the younger people. There is also a significant interaction effect, whereby younger women are more likely to be ex-smokers than their combined age and gender profiles might otherwise suggest.

Table 4 shows two different statistical models that could both be used to describe this data. The two models are statistically equivalent (see for example their identical log likelihoods). Nevertheless, the way in which the effects of the two categorical variables are reported varies between the two models. Model 3.1 is the more conventional presentation, which at first sight better fits the description above. However, Model 3.1 is problematic as a statement about the relative influences of the two explanatory variables, because of some ambiguity over its reference category. It is often forgotten that coefficients and standard errors in a model with interaction terms cannot be readily interpreted independently of each other, since any given coefficient refers to the combined influence of all of the other contributing variables (e.g. Jaccard, 2001: 20).

The more appropriate strategy for describing the interactions between two categorical variables involves specifying a discrete categorical variable that has a distinct value for each combination of circumstances. This is the format used in Model 3.2 (see Table 4). This form of presentation allows the independent effect of each category to be much more easily interpreted. For example, the coefficient for men aged 16–19 in Model 3.2 is –4.29, which means that the chances of younger men reporting being an ex-smoker are significantly lower than those of the reference category (men aged 60–69). Equally, the coefficient for women aged 16–19 is –2.58, which also means that the chances of younger women reporting being an ex-smoker are significantly lower than those of the reference category (men aged 60–69). The presentation of these two parameter estimates in Model 3.2 leads to an easier comparison between the relative chances of young men and young women reporting being an ex-smoker than the parameter estimates presented in Model 3.1. Because the magnitude of the coefficient for younger men is greater than for younger women, we can see that it is younger men who have relatively lower chances of reporting being an ex-smoker.

Table 4 Logistic regression model predicting probability that respondent is an ex-smoker (Parameter estimates for models 3.1 and 3.2)

	Model 3.1			Model 3.2			
	Beta	S.E.	Prob	Beta	S.E.	Prob	Quasi-variance
Male	–	–	–				
Female	–0.65	0.19	.001				
Group 0 (aged 60–69 yrs)	–	–	–				
Group 1 (aged 20–59 yrs)	–0.94	0.15	<.001				
Group 2 (aged 16–19 yrs)	–4.29	1.01	<.001				
Male and Age 60–69				–	–	–	0.017
Male and Age 20–59				–0.94	0.15	<.001	0.005
Male and Age 16–19				–4.29	1.01	<.001	1.009
Female and Age 60–69				–0.65	0.19	<.001	0.020
Female and Age 20–59	0.52	0.22	0.016	–1.08	0.15	<.001	0.005
Female and Age 16–19	2.36	1.10	0.033	–2.58	0.44	<.001	0.175
Constant	–0.44	0.13	.001	–0.44	0.13	.001	
Log likelihood	1684.2 (Pseudo $R^2 = 0.04$)			1684.2 (Pseudo $R^2 = 0.04$)			

n = 3,507

Source: 2002 UK General Household Survey, all adults aged 16–69, unweighted.

When output on categorical interaction effects has been arranged in the format of Model 3.2, quasi-variances for the discrete categories may be calculated in exactly the same way as they would be for a single categorical factor. Again, quasi-variances provide a reliable way of interpreting pairs of contrasts between different combinations of circumstances, which would not have been easily available in a conventional presentation of parameter estimates and standard errors (as in Model 3.1). The quasi-variances reported in Table 4 for Model 3.2 can be used in the manner described above to allow the secondary analyst to rapidly test the significance of contrasts between any two discrete categories.[19] Indeed, whilst Table 4 illustrates a two-way interaction between two categorical explanatory variables, these issues extend readily to higher-order categorical interactions and to interactions between categorical and metric variables (Firth and de Menezes, 2004: 79).

Further Issues in Quasi-variance Statistics

A further concern that we wish to draw attention to is understanding category effects from skewed multiple categorical measures (i.e. variables where large

numbers of cases are concentrated in some categories, and few cases fall into other categories). The most extreme problems concern the situation when the reference category itself is disproportionately sparse. In our experience in this situation it is common that all of the other parameter estimates appear 'insignificant' in relation to the reference category, despite the possible existence of significant contrasts between them. Fortunately, we observe that sociologists most often choose the largest category to be the reference category, whether for a clear substantive reason, or simply because it is often intuitive to consider other groups in relation to the majority group. Nevertheless, we conclude that any situation where the distribution of cases between categories is uneven is likely to increase the chances of misleading interpretations of differences between categories.

An additional appeal of reporting quasi-variances that we have not illustrated in the examples above is the ability to compare statistical models which have different reference categories. Returning to Example 1, consider two models that include Government Office Region as an explanatory variable. In one model the North East region is the reference category and in the other Inner London is the reference category. A reader may wish to understand the effect of living in the North West region. However, in the first model the estimate for the North West region is a comparison with the North East region and in the second model the estimate for the North West is a comparison with Inner London. Again without access to the variance-covariance matrix of parameter estimates a comparison of the effects of living in the North West region in these two analyses cannot be derived, but would be possible if quasi-variances were reported alongside the parameter estimates.

Conclusions

Statistical models provide enormous analytical potential in sociological analyses of survey data. They have been widely deployed across the discipline, and frequently include multiple category explanatory variables. This article has discussed the 'reference category problem', which affects the comparison of categories where one level is not the base category. This problem is not acknowledged in many statistical modelling texts aimed at social researchers (e.g. Cramer, 2003; De Vaus, 2002). Even those treatments (e.g. Hardy, 1993; Hardy and Reynolds, 2004) which illustrate some awareness of this issue have not described solutions that are open to the secondary analyst, such as the reader of published outputs.

We conclude that the quasi-variance calculations described by Firth offer an attractive solution to the reference category problem that can be operationalized by sociological researchers. This is because in standard software information from the variance-covariance matrix of the parameter estimates can be extracted.[20] This information can then be plugged into Firth's web-based calculator and quasi-variances can be estimated. These may be readily used to compare parameter estimates in a manner that will not be influenced by the choice of reference category.

Therefore we are advocating that when sociological researchers estimate models with multiple category explanatory variables they use Firth's web-based calculator to compute quasi-variances and present them alongside usual results such as parameter estimates and their standard errors. The cost of this is simply to add one extra column to tables of results, but the benefit is that the reader of published results is able to reliably make any contrast that they desire. We hope that this article will have raised the general level of awareness of the reference category problem and that the examples have highlighted the benefits of Firth's quasi-variances to the wider sociological community.

Acknowledgements

We thank David Firth and Mick Green for assistance in the development of this paper, Hannah Buchanan-Smith, Saffron Karlson, Richard Lampard and Alison Smith for useful comments on draft versions and the anonymous referees.

Notes

1 See for instance the resources available in the UK Data Archive (www.data-archive.ac.uk/) and the ESDS data support service (www.esds.ac.uk). These resources are emblematic of the growing number of survey datasets that have become available to social scientists and which exhibit high standards of data collection and documentation. In addition, protocols have been established for communicating data quality issues to the reader of research outputs (e.g. Dale, 2006).

2 See the Economic and Social Research Council Postgraduate Training Guidelines (ESRC, 2005: 88).

3 To many readers, multiple and logistic regression techniques will be well-known examples of statistical models. To statisticians, these techniques are two simple examples from the wide class of models recently termed 'Generalised Linear Mixed Models' (GLMM) (e.g. Hedeker, 2005). The reference category problem, the subject of this article, is of relevance to all examples of GLMMs (Firth, 2003).

4 Sociologists may be acquainted with two situations where the reference category effect is not reported as zero. In both instances, the original model estimation still sets the reference category coefficient to equal zero, but, in order to aid communication, a further statistical transformation is applied which alters the values of the coefficient effects as they are reported. The first example refers to the case of categorical regression models (e.g. logistic regression), when analysts often report odds ratios, rather than the value of the coefficient estimates. In this instance, the odds ratios are calculated as a logarithmic function of the original coefficient estimates, which means that a reference category coefficient value of zero generates an odds ratio of one (i.e. the exponential of zero). For this reason, the coefficient effect of the reference category is often reported as one in outputs which display odds ratios. The second situation is when an alternative form of coding is reported. The most usual strategy is 'indicator coding', which involves reporting the reference category effect as zero, but a possible

alternative is 'deviation coding', which involves ensuring that the reported parameter estimates are calculated in such a way that they sum to the value one (meaning that the reference category parameter is reported as the value one minus the sum of all other parameters). Such alternative coding transformations are readily obtained from model estimations using SPSS, and may also be obtained, after some programming, using Stata. Alternative coding strategies have no impact upon the 'reference category problem' under discussion here, and our examples concentrate upon the most commonly employed strategy of 'indicator coding'.

5 There is no strict protocol for reporting the estimates of statistical models in sociological analysis, although there are conventions. We observe that it is common for many sociologists to report parameter estimates (which may also be referred to as betas, coefficients or estimates). Alongside parameter estimates, standard errors are often reported. Sociologists will commonly report associated p values (or probabilities) or indicate significance at a certain level (e.g. $p<0.05$). Other analysts will provide confidence intervals for parameter estimates, calculated directly from the standard errors. In all cases, it is important to understand that these estimates relate to the contrast between the category of interest and the reference category.

6 The SARs may be downloaded after registration with the UK Census Registration Service, http://census.data-archive.ac.uk/. The full GHS data may be accessed from the UK Data Archive, www.data-archive.ac.uk/, although an extract file used in the worked examples is freely downloadable from our website (www.longitudinal.stir.ac.uk/qv). This data file is also used by Fielding and Gilbert (2006).

7 Hardy and Reynolds (2004) also note that a common shortcut to undertaking these formal tests involves the primary analyst simply repeating the model with a variety of alternative choices of reference category – therefore building up a series of all possible contrasts (to the reference category). This can prove a helpful strategy, but again it is not available to a secondary analyst such as the reader of published output. Moreover, the primary analyst will need to make a choice over which level of the variable they ultimately present as their reference category in published work.

8 Examples on our website illustrate how the appropriate data can ultimately be obtained in SPSS. We thank Mick Green, Lancaster University, for suggestions for obtaining covariance values from SPSS.

9 Many statistical packages, such as Stata, have pre-programmed routines for undertaking particular comparisons on a wide range of alternative categories (some illustrations of Stata procedures are on our webpages). However, such facilities are not currently available in SPSS.

10 In a model with q parameters there would, in general, be $\frac{1}{2}q\,(q-1)$ covariances to report. Therefore, reporting the matrix is seldom, if ever, feasible in paper-based publications. However, following the recommendation made in Dale (2006), internet sites could be used to publish large matrices.

11 We refer to this as Firth's method but are aware that he notes that the initial suggestion that quasi-variance statistics may be of value was made by Ridout (1989).

12 Quasi-variances are generic statistics, which may readily be calculated for categorical variable estimates associated with almost any form of statistical model (Firth and de Menezes, 2004). Firth (2003) illustrates this generality by applying the method to two specialist sociological statistical applications, an advanced log-linear model and a multinomial logit model.

13 Firth has also provided program routines to generate quasi-variance statistics using some other specialist statistical packages (see Firth, 2000, 2006). We are currently investigating the construction of an extension program file in Stata which would allow the calculation of quasi-variances entirely within that package (see www.longitudinal.stir.ac.uk/qv).
14 This information may be supplied in two alternative formats. One method involves entering the lower triangle of the variance-covariance matrix itself (this matrix is readily obtained from Stata, and we recommend this format as the more intuitive). However, the equivalent information may also be supplied through a column of standard errors for each parameter estimate, alongside the lower triangle of the estimates correlation matrix (this format is more accessible for analysts using SPSS, since SPSS does not readily supply the variance-covariance matrix of estimates for all types of model). Our online example files illustrate the derivation of this information, using both SPSS and Stata, for the examples discussed in this article.
15 The calculator also reports quasi-standard errors (i.e. $\sqrt{\text{quasi-variance}}$).
16 www.longitudinal.stir.ac.uk/qv/qv_varest.xls
17 As described above, quasi-variance statistics are approximations that allow us to undertake comparison tests without requiring complex data from the variance-covariance matrix of parameter estimates. The accuracy of these approximations is therefore a question of concern. However, we can suggest that inaccuracy is likely to be negligible for most sociological examples where large-scale secondary survey data are analysed and when relatively well-specified statistical models are employed. This issue is further discussed on our website (www.longitudinal.stir.ac.uk/qv).
18 Readers may note that Table 3, in common with all of the examples in this text, presents unstandardized parameter coefficients. It is possible to calculate quasi-variance statistics for standardized variables, but care must be taken to ensure that data standardization precedes the model estimation. We recommend the use of unstandardized variable estimations as a simple way to avoid confusion in this issue.
19 For example, the contrast between the two male categories for age 20–59 and 16–19 yields a Wald test statistic of 11.07 at one degree of freedom, which indicates a significant difference in the coefficient values for the two categories (namely, younger men are significantly less likely to report being an ex-smoker than men in the medium age category).
20 As we have noted above this is more straightforward in Stata than in SPSS.

References

Allison, P.D. (1999) *Multiple Regression: A Primer*. London: Sage.
Berk, R.A. (2004) *Regression Analysis: A Constructive Critique*. London: Sage.
Cramer, D. (2003) *Advanced Quantitative Data Analysis*. Maidenhead: Open University Press.
Dale, A. (2006) 'Quality Issues with Survey Research', *International Journal of Social Research Methodology* 9(2): 143–58.

Dale, A. and R.B. Davies (eds) (1994) *Analysing Social and Political Change: A Casebook of Methods*. London: Sage.
De Vaus, D. (2002) *Analyzing Social Science Data: 50 Key Problems in Data Analysis*. London: Sage.
ESRC (2005) *Economic and Social Research Council Postgraduate Training Guidelines*, 4th edn. Swindon: Economic and Social Research Council.
Fielding, J.L. and G.N. Gilbert (2006) *Understanding Social Statistics*, 2nd edn. London: Sage.
Firth, D. (2000) 'Quasi-variances in Xlisp-Stat and on the Web', *Journal of Statistical Software* 5(4): 1–13.
Firth, D. (2003) 'Overcoming the Reference Category Problem in the Presentation of Statistical Models', *Sociological Methodology* 33(1): 1–18.
Firth, D. (2006) *qvcalc: Quasi-variances for Factor Effects in Statistical Models (R package version 0.8–3)*. Vienna: R Foundation for Statistical Computing (http://www.warwick.ac.uk/go/qvcalc).
Firth, D. and R.X. de Menezes (2004) 'Quasi-variances', *Biometrika* 91(1): 65–80.
Goldthorpe, J.H. (2007) *On Sociology: Numbers, Narratives, and the Integration of Research and Theory* (2nd Edition). Stanford: Stanford University Press.
Hardy, M. (1993) *Regression with Dummy Variables*. London: Sage.
Hardy, M. and J. Reynolds (2004) 'Incorporating Categorical Information into Regression Models: The Utility of Dummy Variables', in M. Hardy and A. Bryman (eds) *Handbook of Data Analysis*, pp. 209–36. London: Sage.
Hedeker, D. (2005) 'Generalized Linear Mixed Models', in B. Everitt and D.C. Howell (eds) *Encyclopaedia of Statistics in Behavioural Science*, pp. 729–38, URL (consulted October 2007): http://mrw.interscience.wiley.com/emrw/9780470013199/esbs/article/bsa251/current/abstract. New York: Wiley.
Jaccard, J. (2001) *Interaction Effects in Multiple Regression*. London: Sage.
Menard, S. (2001) *Applied Logistic Regression Analysis*, 2nd edn. London: Sage.
ONS (2005) *2001 Individual Sample of Anonymised Records (Licensed File) [computer file]*. London: Office for National Statistics, Census Division, [original data producer(s)]. University of Manchester, Cathie Marsh Centre for Census and Survey Research [distributor], August.
ONS Social Survey Division (2006) *General Household Survey, 2001–2002 [computer file]*, 4th edn. Colchester: UK Data Archive [distributor], February. SN: 4646.
Ridout, M.S. (1989) 'Summarizing the Results of Fitting Generalized Linear Models to Data from Designed Experiments', in A. Decarli, B. Francis, R. Gilchrist and G. Seeber *Statistical Modelling: Proceedings of GLIM89 and the 4th International Workshop on Statistical Modelling*, pp. 262–9. New York: Springer-Verlag.
Rose, D. and D.J. Pevalin (2003) *A Researcher's Guide to the National Statistics Socio-Economic Classification*. London: Sage.
SPSS (2005) *SPSS for Windows, Release 14.0.0*. Chicago, IL: SPSS Inc.
Stata (2005) *Intercooled Stata 9.0 for Windows*. College Station, TX: StataCorp LP.

Vernon Gayle

Is Senior Lecturer in Sociology at the University of Stirling. His research expertise includes the analysis of social surveys, particularly longitudinal data, with an empirical interest in youth and social divisions.
Address: Applied Social Science, University of Stirling, Stirling FK9 4LA, UK.
E-mail: vernon.gayle@stirling.ac.uk

Paul Lambert

Is Lecturer in Sociology at the University of Stirling. His research expertise includes the analysis of occupational data and ethnicity through secondary surveys.
Address: Applied Social Science, University of Stirling, Stirling FK9 4LA, UK.
E-mail: paul.lambert@stirling.ac.uk