Using Quasi-variance to Communicate Sociological Results from Statistical Models

Working Paper 3 of 'Longitudinal Data Analysis for Social Science Researchers',

ESRC Researcher Development Initiative training programme

http://www.longitudinal.stir.ac.uk/

Dr Vernon Gayle
Dr Paul S. Lambert

Applied Social Science
University of Stirling
vernon.gayle@stirling.ac.uk

Last updated 30.8.06.

This paper is currently under review for publication.

Introduction

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 (Firth 2000; 2003; Firth and 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 paper describes Firth's idea of 'quasi-variance' as a solution to the 'reference category problem' in presenting outputs from statistical models. We provide a number of examples to illustrate the flexibility of the quasi-variance approach, and discuss 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 packages SPSS and Stata, to help sociological researchers practice and apply this technique. We conclude that quasi-variances should be routinely presented as output when the results of statistical models with categorical explanatory variables are reported within sociological analyses.

Survey Analysis and Statistical Models in Sociological Research

Evaluations of variable analyses in sociology date back at least fifty years (see Blumer 1956). Over the decades a virtual industry producing critiques of variable analyses from various standpoints has developed. We suggest that arguments for and against variable analysis, and in particular the analysis of data from social surveys, have at times resembled a caricature not dissimilar to the Shakespearean feud between the Montagues and the Capulets. In this paper we do not wish to either visit or reopen these debates. However, we would like to note a comment made by Goldthorpe that critics of survey based sociological research ritually characterise it as static and this is simply to ignore the rapid development of survey related work (Goldthorpe 2000 p.17).

We would like to emphasise that, far from remaining static in their practices, survey researchers in sociology have been reflexive and have attempted to respond to various critiques and specific shortcomings that have been highlighted. These responses have usually involved improving statistical techniques of variable analysis, and improving the underlying data quality. A large number of new statistical methods appropriate for the analysis of social survey data have emerged in recent decades[1]. Increasingly, these methods have been integrated into mainstream statistical software packages (e.g. SPSS 2004; STATA 2005) and are now widely available to sociological researchers. At the same time a growing number of surveys and datasets have become available to social scientists, which exhibit high standards of data collection and documentation[2],

[1] Informative recent collections which summarise emerging statistical methods include Dale and Davies (1994); Chambers and Skinner (2003); Harkness *et al* (2003); Hardy and Bryman (2004); and Skrondal and Rabe-Hesketh (2004).

[2] See for instance the resources available in the UK Data Archive (http://www.data-archive.ac.uk/) and the ESDS data support service (http://www.esds.ac.uk).

and well-established protocols exist for communicating data quality issues to the reader of research outputs (e.g. Dale 2006).

Statistical models offer an attractive way for sociological researchers to summarise patterns from social survey datasets (Dale and Davies 1994; Goldthorpe 2000). They offer techniques to summarise 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 parameters 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 emphasise (Berk 2004).

This paper concentrates on the process of communicating statistical models in order to describe the relative effects associated with multiple category explanatory variables. Many explanatory variables in social science research are categorical, by which we mean they are measured according to membership of a number of discrete categories[4]. Almost all standard statistical models can readily incorporate categorical explanatory variables in their specification[5]. However Firth's papers (2000, 2003) highlight a limitation in standard practices for communicating parameter estimates from categorical explanatory variables, related to how the effects of different categories of the same explanatory variable are reported and interpreted.

The Reference Category Problem

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 benchmark effect is arbitrarily fixed to zero[6], and other category effects are interpreted as the additional impact of not being

[3] To many readers, multiple and logistic regression techniques will be well-known examples of statistical models. To statisticians, these techniques are two examples from the wider class of models often termed as 'Generalised Linear Models' (e.g. Nelder and Wedderburn 1972; Hedeker 2005).

[4] The following list indicates that statistical models with categorical explanatory variables are found in papers presented in sociological journals in a wide variety of substantive areas. Connolly (2006) in *British Educational Research Journal;* Harsløf (2005) in *Journal of Youth Studies;* Pahl and Pevalin (2005) in *The British Journal of Sociology;* Van de Werfhorst (2005) in *Acta Sociologica;* Sandu (2005) in *Current Sociology;* Widmer, Kellerhals and Levy (2004) in *European Sociological Review.*

[5] The standard estimation strategy involves creating 'dummy' or 'indicator' variables, where a suite of variables describes membership (or not) of each discrete category, for an extended discussion see Hardy (1993).

[6] The numeric specification of the benchmark effect is sometimes adjusted by using alternative types of coding. These alternatives are readily implemented in SPSS but are also possible, with some

in the benchmark category. Standard statistical software undertakes formal comparisons of whether or not each category effect differs from the benchmark 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 analyst, such a those reading published results, cannot make such comparisons themselves[7]. As we shall describe, 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), the General Household Survey (GHS) of 2002, and a special example of a panel survey dataset. We have chosen these datasets because they can be freely downloaded[8]. 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 can be downloaded from our website (www.longitudinal.stir.ac.uk/qv/)[9].

Example 1 and an Introduction to Quasi-variance

The first example (model 1, shown in 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 twelve months (0= no; 1= yes). There are

---

programming, in Stata. The most common strategy is 'indicator coding', which involves forcing the base category effect to equal zero. A notable alternative is 'deviation coding', which involves ensuring that the final parameter estimates are constrained to sum to the value one. 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'.

[7] 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. $p<.05$). 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.

[8] 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, http://www.data-archive.ac.uk/, although an extract file used in the worked examples is freely downloadable from the website above (this data file is also used by Fielding and Gilbert 2006).

[9] Example 3 is only illustrated through a Stata command file, since SPSS does not have functionality to estimate the relevant panel model with a binary outcome variable.

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. ten). 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 (.09) that is significantly different to the North East region ($p<.001$). Yorkshire and Humberside Region is also significantly different to the North East with an estimate of .12 ($p<.001$).

It is plausible that a reader may wish to make other comparisons between Government Office Regions. For example a researcher 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 (.07 to .11) with the equivalent interval for Yorkshire and Humberside (.10 to .14). This is presented graphically in Figure 1. 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 are not confidence intervals for the reference category because it is forced to equal zero. A useful analogy to conceptualise this is to consider that the confidence intervals for other categories are 'artificially' wider, because the reference category has no confidence interval.

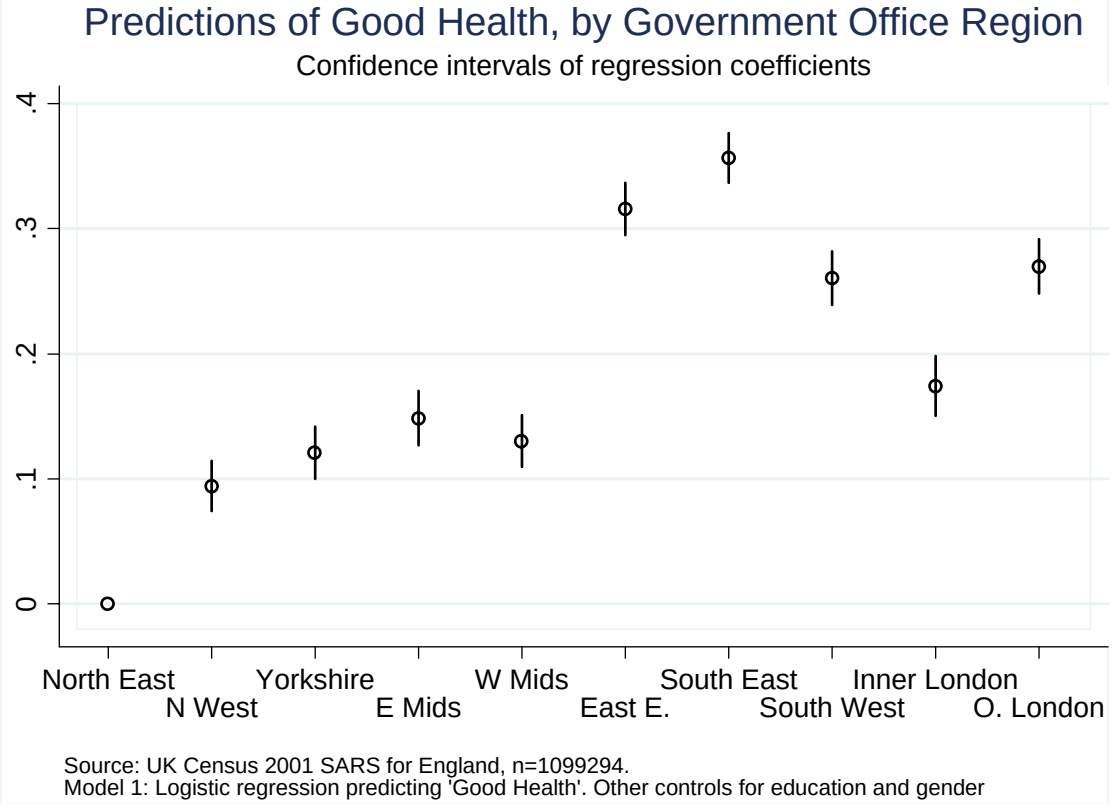
Table 1 Model 1

Logistic regression prediction that self-rated health is 'good'.
(Parameter estimates for model 1)

	Beta	Standard Error	Prob.	95% Confidence Intervals	
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	-	-	-	-	-
North West	0.09	0.0102	<.001	0.07	0.11
Yorkshire & Humberside	0.12	0.0107	<.001	0.10	0.14
East Midlands	0.15	0.0111	<.001	0.13	0.17
West Midlands	0.13	0.0106	<.001	0.11	0.15
East of England	0.32	0.0107	<.001	0.29	0.34
South East	0.36	0.0101	<.001	0.34	0.38
South West	0.26	0.0109	<.001	0.24	0.28
Inner London	0.17	0.0122	<.001	0.15	0.20
Outer London	0.27	0.0111	<.001	0.25	0.29
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.

Figure 1


Predictions of Good Health, by Government Office Region
Confidence intervals of regression coefficients
0
.1
.2
.3
.4
North East
N West
Yorkshire
E Mids
W Mids
East E.
South East
South West
Inner London
O. London
Source: UK Census 2001 SARS for England, n=1099294.
Model 1: Logistic regression predicting 'Good Health'. Other controls for education and gender

Comparing Categories - Conventional Calculations

Continuing with example 1 (above), 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 evaluating a t-statistic for the unstandardised parameter estimates given in equation 1 (for a detailed discussion see Hardy and Reynolds 2004)[10].

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

It is simple enough to compute the difference between the two beta estimates for the North West and Yorkshire and Humberside (.09-.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:

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

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[11]. 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:

$$0.0083 = \sqrt{.00010483 + .00011543 - 2\,(.00007543)}$$

[10] Hardy and Reynolds (2004) also note that a common short-cut to undertaking these formal tests involves the 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 sensible strategy for the primary analyst, but again it is not available to a secondary analyst such as the reader of published output, and moreover the primary analyst will need to make a choice over which level of the variable they ultimately present as their reference category.

[11] 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.

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

$t = -.03 / 0.0083 = -3.2$

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 ($p<.05$). 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:

$Wald\ \chi^2 = (-.03 / 0.0083)^2 = 10.22$; $p = .0014$.

The value of 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 factors of a categorical variable are seldom undertaken and reported in sociological outputs. This occurs despite their obvious substantive value (consider for instance cases, such as in example 1, where the different categories represent areas where groups or organisations may have independent policy making capacities). Moreover, the range of comparisons that may be tested extends beyond simple two-category contrasts. Multi-category contrasts may also be deployed (an example would be to ask whether all of the northern and midland regions of Example 1 are significantly different to all of the southern and eastern regions)[12]. 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 webpage) however this facility is not currently available in SPSS.

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[13]. Firth's (2003) recommendation, that analysts routinely display 'quasi-variance' statistics for all multiple category explanatory variables, offers a neat and practical solution this impasse.

---

[12] Quasi-variance statistics can be used to simplify the estimation of multiple-contrasts, but the mathematical calculations are relatively complex (see Firth and Menezes 2004). Here we concentrate on two-way contrasts.

[13] In a model with q parameters there would, in general, be ½q (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.

Table 2 Variance Covariance Matrix of Parameter Estimates (Model 1)

	Column	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Row		*North West*	*Yorkshire & Humberside*	*East Midlands*	*West Midlands*	*East England*	*South East*	*South West*	*Inner London*	*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*

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[14]. 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[15]. Firth provides an online calculator (see Figure 2) which we use in this illustration[16]:

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

To use the online calculator, the primary 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 ten Government Office Regions). The second is information from the variance covariance matrix of the parameter estimates. This information may be supplied in two alternative formats. It may take the form of the lower triangle of the variance covariance matrix itself (this format is readily obtained from Stata, and we suggest that this format is the more intuitive, and therefore should be the preferred option). 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 allows for the immediate supply of standard errors and the relevant correlation tables, but it does not always readily supply the variance covariance matrix of estimates)[17].

In our experience, the precise format of the necessary data from the variance covariance matrix has confused some colleagues. To help avoid confusion, Figure 3

[14] 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).

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

[16] Firth has also provided programme routines to generate quasi-variance statistics using some other specialist statistical packages (see Firth 2000; 2006).

[17] Our online example files illustrate the derivation of this information for the four examples discussed in this paper. The Stata example files show this by using the lower triangle of the variance covariance matrix of parameter estimates, whereas the SPSS examples illustrate how the information is supplied through the column of standard errors and the appropriate portion of the estimates correlation table.

depicts the information (for Example 1) from the variance covariance matrix of the parameter estimates that must be entered in the data window of the online calculator (ie, this is the format that users of Stata would ordinarily supply).

The web-based calculator produces a quasi-variance for each level of the categorical explanatory variable[18]. For Example 1, the outputs from the quasi-variance estimates are reported in Table 3 (which contains a simple extension to 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:

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

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 =.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[19]. 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.

In practice, we have found that a graphical example has helped researchers to better comprehend this issue. In Figure 4 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. For better illustration we have plotted them alongside parameter estimates and conventional confidence intervals[20]. In Figure 4 we can see that using quasi-variances, the confidence intervals for the North West and Yorkshire and Humberside no longer overlap. A useful analogy to help understand this difference is to imagine that the conventional confidence intervals are larger, because they are related to estimates that are compared to the based category that has to be fixed to zero.

[18] The calculator also reports quasi-standard errors (i.e. $\sqrt{\text{quasi - variance}}$).

[19] http://www.longitudinal.stir.ac.uk/qv/qv_varest.xls

[20] Constructed from beta ± (1.96 * standard error).

Figure 2 David Firth's Web Based Quasi-Variance Calculator

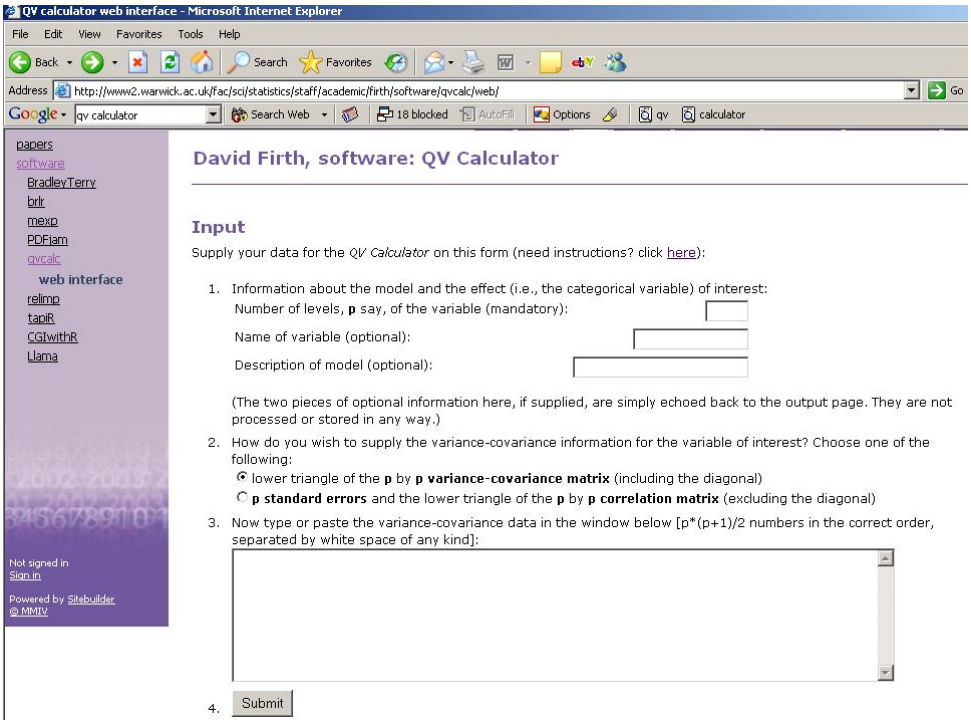


Figure 3 Information from the Variance-Covariance Matrix Entered into the Data Window (Model 1)

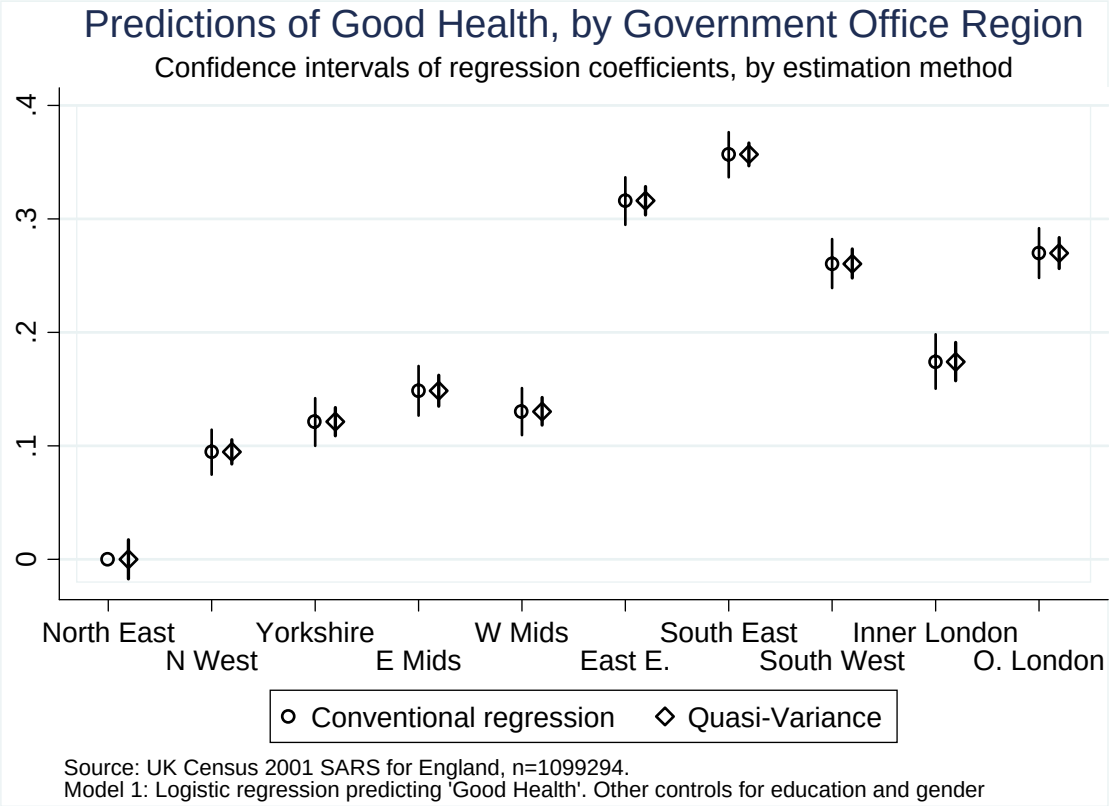
0									
0	0.00010483								
0	0.00007543	0.00011543							
0	0.00007543	0.00007543	0.00012312						
0	0.00007543	0.00007543	0.00007543	0.00011337					
0	0.00007544	0.00007543	0.00007543	0.00007543	0.00011480				
0	0.00007545	0.00007544	0.00007544	0.00007544	0.00007545	0.00010268			
0	0.00007544	0.00007543	0.00007544	0.00007543	0.00007544	0.00007546	0.00011802		
0	0.00007552	0.00007548	0.00007550	0.00007547	0.00007554	0.00007572	0.00007558	0.00015002	
0	0.00007547	0.00007545	0.00007546	0.00007545	0.00007548	0.00007555	0.00007549	0.00007598	0.00012356

Table 3

Logistic regression prediction that self-rated health is 'good'.
(Parameter estimates for regional variable from model 1, and Quasi-variance statistics – compare with Table 1)

	Beta	Conventional Standard Error	Quasi-variance
North East	0.00	-	0.0000755
North West	0.09	0.0102	0.0000294
Yorkshire & Humberside	0.12	0.0107	0.0000400
East Midlands	0.15	0.0111	0.0000477
West Midlands	0.13	0.0106	0.0000380
East of England	0.32	0.0107	0.0000394
South East	0.36	0.0101	0.0000272
South West	0.26	0.0109	0.0000426
Inner London	0.17	0.0122	0.0000743
Outer London	0.27	0.0111	0.0000480

Figure 4


Predictions of Good Health, by Government Office Region
Confidence intervals of regression coefficients, by estimation method
0
.1
.2
.3
.4
North East
N West
Yorkshire
E Mids
W Mids
East E.
South East
South West
Inner London
O. London
Conventional regression
Quasi-Variance
Source: UK Census 2001 SARS for England, n=1099294.
Model 1: Logistic regression predicting 'Good Health'. Other controls for education and gender

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 data on the age at which the individual left full-time education (min=10; max=50; mean=17.35; s.e.=.05). The model includes explanatory variables for age and social class. For the latter 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 4 gives a conventional presentation of the results from this model[21], with the additional presentation of quasi-variance statistics (illustrations of the computation of these statistics are on our webpage).

Table 4

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

	Beta	**Standard Error**	**Prob.**	**Quasi-variance**
Age in years – 40	-0.053	0.003	<0.001	-
Social class:				
Advantaged (n=1679)	-	-	-	0.0038
Lower-supervisory (n=279)	-1.798	0.163	<.001	0.0228
Semi-Routine (n=524)	-1.931	0.126	<.001	0.0121
Routine (n=397)	-2.325	0.141	<.001	0.0160
Constant	18.403	0.063	<.001	-

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

The reference category for the social class variable in model 2 is 'advantaged' occupations (although, as indicated in Table 4, our definition of 'advantaged' includes the majority of respondents). One interesting question would be whether educational experiences for those in the 'lower supervisory' classification are significantly different to those from 'routine occupations'.

[21] Readers may note that Table 4, in common with all of the examples in this text, presents unstandardised parameter coefficients. The presentation of standardised coefficients is sometimes used in areas of sociological research, and can add some useful additional information to an output table. For example standardised coefficients may be used to make statements about the relative importance of different explanatory factors from different variables. However the calculation of quasi-variance statistics requires unstandardised coefficients in combination with unstandardised variance-covariance information.

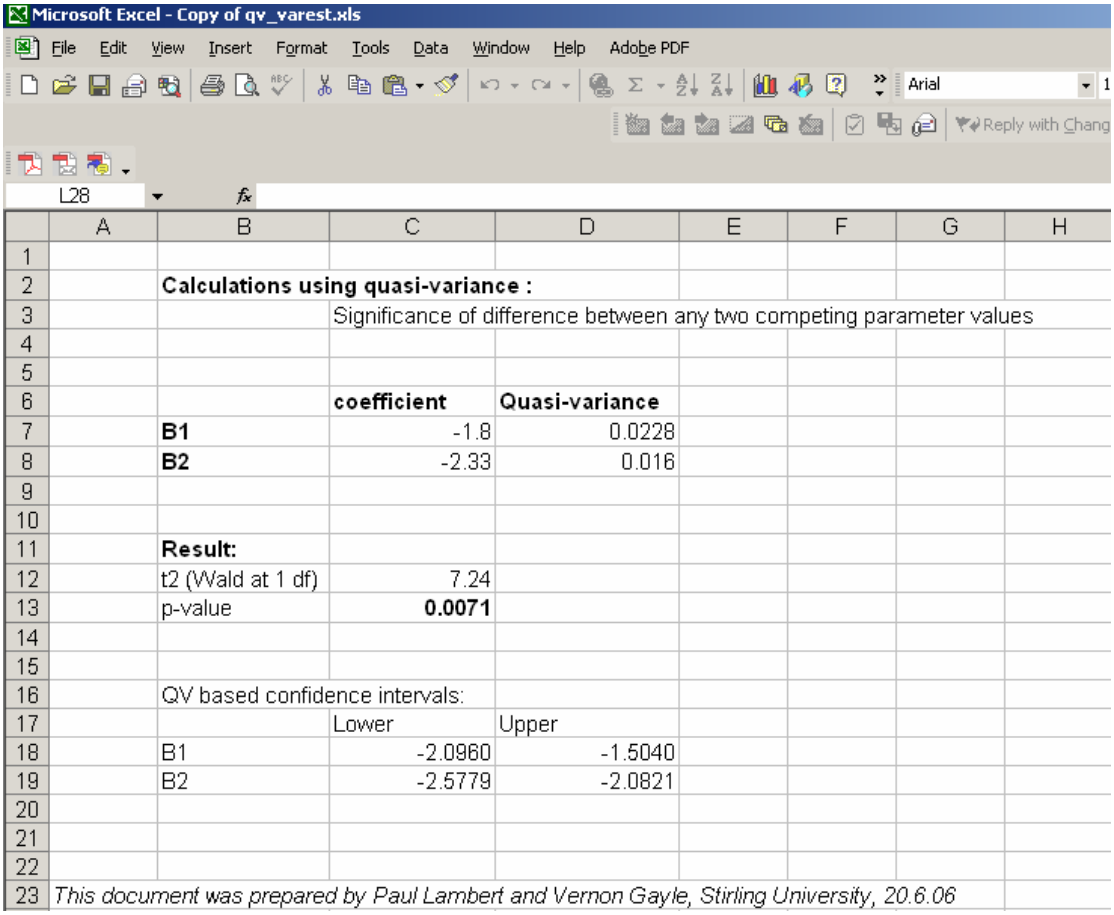
Because quasi-variances have been reported, this test can be readily conducted on the basis of the Table 4 outputs. Using the same notations as Example 1:

$$t = \frac{\hat{\beta}_2 - \hat{\beta}_4}{\sqrt{quasi\ \mathrm{var}(\hat{\beta}_2) + quasi\ \mathrm{var}(\hat{\beta}_4)}} = \frac{-1.80 + 2.33}{\sqrt{0.0228 + 0.016}}$$

$$Wald\ \chi^2 = 7.24\ ;\ p = 0.007\ .$$

This calculation could be undertaken by using the excel calculator available from our website (Figure 5). Therefore 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.

Figure 5



	A	B	C	D	E	F	G	H
1								
2		**Calculations using quasi-variance :**						
3			Significance of difference between any two competing parameter values					
4								
5								
6			**coefficient**	**Quasi-variance**				
7		**B1**	-1.8	0.0228				
8		**B2**	-2.33	0.016				
9								
10								
11		**Result:**						
12		t2 (Wald at 1 df)	7.24					
13		p-value	**0.0071**					
14								
15								
16		QV based confidence intervals:						
17			Lower	Upper				
18		B1	-2.0960	-1.5040				
19		B2	-2.5779	-2.0821				
20								
21								
22								
23	*This document was prepared by Paul Lambert and Vernon Gayle, Stirling University, 20.6.06*							

Example 3

To further illustrate the flexibility of quasi-variance methods and their propriety across a range of statistical models, in our next example we fit a panel model. There are a large number of longitudinal datasets with panel elements that are appropriate for sociological analysis[22]. Panel models are well established within economics but are generally less well known in sociology[23]. In this example we fit a random effects logit model to synthetic data that mirrors records collected in a 1988 survey for the ESRC funded Social Change and Economic Life Initiative (Gaille 1988).

The data is a small panel of women who are observed yearly for up to 14 years. The outcome variable is the woman's employment status (0=not working; 1= employed). The model includes a variable for the husband's employment status (0=employed; 1=unemployed), a variable related to childcare (0=no children under 1 year; 1=children under 1 year) and a categorical variable representing age bands (3= 31+ years; 2= 26-30 years; 1= 21-25 years; 0= 18-20 years).

The results of the model are reported in Table 5. Once again quasi-variances are reported alongside the estimates for the multiple category explanatory variable. As an example, we are interested in comparing those women who are aged 26-30 years with counterparts who are aged 21-25 years. Again this comparison is relatively straightforward and simply requires the estimates and the quasi-variances to be plugged into our web-based calculator.

$$t = \frac{1.09 - 0.42}{\sqrt{0.0670 + 0.0239}}$$

Wald χ^2 *=4.94*; p =.026 .

Therefore in this panel model of women's employment there is a significant difference between women aged 26-30 years and woman aged 21-25 years.

Model 3 also provides an interesting example of a categorical variable with an uneven (or 'skewed') distribution of cases in each category. As shown in Table 5, only 13 women occupied the youngest age band at some point in time over the period that the panel were observed. This is low compared with the number of women in the other age bands. Categorical explanatory variables with uneven distributions are common in social surveys. A common example being measures of ethnicity where even in nationally representative surveys, some groups will have low numbers. Skewed categorical explanatory variables raise an important issue for interpretation of the differences between categories, and in practice this is where the reference category problem tends to be most dramatic.

For instance, in Model 3, the age bands variable is significant (with a change in deviance of 10 at 3 degrees of freedom), and there are significant contrasts between several of the categories of the age bands variable, such as between the 21-25 and 26-30 age category. In fact, these effects are not apparent from conventional presentation,

[22] For a flavour of these readers who are new to this area might consult http://www.statistics.gov.uk/downloads/theme_compendia/Tracking_v8.pdf

[23] For a good review see Halaby (2004).

where all coefficient estimates for the age band effects were non-significant at a 95% threshold when contrasted to the reference category. In this example, the reason that the conventional presentation is misleading is because the explanatory variable is skewed.

We observe that in many statistical models, sociologists choose the largest category to be the reference category. This might be either for a good substantive reason, or simply because it is often intuitive to consider other groups in relation to the majority group. The point to remember here is that the parameter estimates for the other categories of the explanatory variable are estimated in relation to this reference category.

In other examples we observe that analysts have set a small group as the reference category. Again the parameter estimates for the other categories of the explanatory variables are estimated in relation to this category. In practice this can have an even more dramatic effect than having a large group as the reference category when attempting to compare estimates of the other levels of the explanatory variable. For this reason we argue that special care should be taken when comparing categories of unevenly distributed categorical variables. Ultimately genuine comparisons cannot be made without access to the variance-covariance matrix of parameter estimates, but once again the presentation of quasi-variances offers an effective solution.

Table 5

Random effects logit model predicting probability of married women's employment.
(Parameter estimates for model 3)

	Beta	Standard Error	Prob.	Quasi-variance
Husband's employment status:				
Employed	-	-	-	-
Unemployed	-2.30	0.44	<0.001	-
Number of children in house:				
No children under 1 year	-	-	-	-
Children under 1 year	-2.35	0.34	<0.001	-
Woman's age:				
18-20 years (n=13)	-	-	-	0.554
21-25 years (n=61)	1.09	0.76	.151	0.067
26-30 years (n=80)	0.42	0.77	.584	0.024
30+ years (n=117)	1.42	0.78	.069	0.036
Constant	0.11	0.78	.882	-

n= 155 (observations = 1580)
Log likelihood = -660.7
Source: 'wemp.dat' synthetic data file (www.longitudinal.stir.ac.uk/qv).

Example 4

Interaction effects in statistical models are often substantively important to analyses but in our experience they can be tricky to report and, frequently the effects of interactions are 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 6 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 4.1 and 4.2 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 6 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 4.1 is the more conventional presentation, which at first sight better fits the description above. However, Model 4.1 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 of 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 and Turrusi 2003, p20).

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 4.2 of Table 6. 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 4.2 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 age 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). Moreover, the presentation of these two parameter coefficients in Model 4.2 leads to an easier comparison between the relative chances of young men and young women reporting being a ex-smoker than those of Model 4.1 permit. 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.

When categorical interaction data has been arranged in the format of model 4.2, quasi-variances and quasi-standard errors for the discrete categories may be calculated in exactly the same way as they would be for a single categorical factor. Again, such quasi-variances provide a reliable way of interpreting pairs of contrasts between different combinations of circumstances, which would not have been available in a conventional presentation of parameter estimates and standard errors (as in Model 4.1). Thus, the quasi-variances reported in Table 6 for model 4.2 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[24]. Indeed, whilst Table 6 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 Menezes 2004, p79).

Table 6

Logistic regression model predicting probability that respondent is an ex-smoker.
(Parameter estimates for models 4.1 and 4.2)

	Model 4.1			Model 4.2			
	Beta	**S.E.**	**Prob**	**Beta**	**S.E.**	**Prob**	**Quasi-variance**
Male	-	-	-				
Female	-0.65	0.19	0.001				
Group 0 (60-69yrs)	-	-	-				
Group 1 (20-59yrs)	-0.94	0.15	<0.001				
Group 2 (16-19yrs)	-4.29	1.01	<0.001				
Male and Age 60-69				-	-	-	0.017
Male and Age 20-59				-0.94	0.15	<0.001	0.005
Male and Age 16-19				-4.29	1.01	<0.001	1.009
Female and Age 60-69				-0.65	0.19	<0.001	0.020
Female and Age 20-59	0.52	0.22	0.016	-1.08	0.15	<0.001	0.005
Female and Age 16-19	2.36	1.10	0.033	-2.58	0.44	<0.001	0.175
Constant	-0.44	0.13	0.001	-0.44	0.13	0.001	
Log-likelihood	1684.2 (Pseudo R^2=0.04)			1684.2 (Pseudo R^2=0.04)			

n = 3507
Source: 2002 UK General Household Survey, all adults aged 16-69, unweighted.

[24] 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).

Further issues in quasi-variance statistics

In an outstanding contribution to the sociological understanding of another specialist methodological issue in the use of statistical models, Stolzenberg and Relles (1997) seek to convey to research oriented analysts the circumstances where 'selection modelling' techniques are likely to be most relevant to sociological applications. As indicated in our discussion of Examples 3 and 4 above, there are two circumstances where, similarly, we suggest that attention to the reference category problem is particularly important to sociological researchers. The first (e.g. Example 3) concerns 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 which case it is common that all other parameter estimates appear 'insignificant', despite the possible existence of significant contrasts within them). However, in our experience, any situation where the distribution of cases between categories is uneven is likely to increase the chances of misleading interpretations of differences between categories.

The second situation where we suggest that reference category problems are greater concerns understanding interaction effects between different categorical variables (e.g. Example 4). In these circumstances, common strategies for reporting interaction effects (as for instance Model 4.1 in Table 6) can be misleading about the different relative contrasts between two variables. Part of the solution is simply a more thoughtful arrangement of the explanatory variable effects (e.g. Model 4.2 in Table 6). However the presentation of quasi-variance statistics makes a helpful contribution in clarifying the nature of the multiple categorical effects and allowing tests for the differences between groups.

An additional appeal of reporting quasi-variances that we have not illustrated in the examples above is the ability to compare model results in different studies. In the most obvious sense this might be where a statistical model is parameterised differently in two studies. Returning to example 1, consider two studies that report a model that includes Government Office Region as an explanatory variable. In one study 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.

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. Firth and Menezes (2004) explore this accuracy in some detail (see also Menezes 1999). Firstly, for the specific case of a multiple category explanatory variable with three categories, it is reported that quasi-variances

are (necessarily) exactly accurate. For other cases, the level of accuracy of quasi-variance statistics may be readily calculated from the same information needed to generate them. Firth's web calculator undertakes this calculation automatically, by invoking a specialised programme using the R package (Firth 2006). The accuracy calculations generated by the web calculator show the range between the highest and lowest levels of mismatch between quasi-variance based conclusions and the results of the conventional test comparison (i.e. the differences between using expressions (2) and (3))[25]. Two sets of ranges are generated, showing the accuracy of all possible two-way comparisons, and those of all possible comparisons (including multi-way comparisons). The former are usually of most interest. Broadly, for larger survey samples, quasi-variance approximations tend to be extremely accurate. For instance the largest inaccuracy in any of our examples 1-4 was 2%, meaning that the difference in significance calculations between the two methods was always highly accurate[26]. Indeed, 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. Nevertheless it is worth suggesting that researchers should review the range of inaccuracies for each quasi-variance set calculated, and report instances where an inaccuracy exceeds 10% (c.f. Firth 2003 p8).

Conclusions

Statistical models provide enormous analytical potential in sociological analyses of survey data. As we have argued they have been widely deployed across the discipline, and frequently include multiple category explanatory variables. This paper 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 of the introductory texts on statistical modelling techniques that are targeted at social researchers (e.g. de Vaus 2002; Cramer 2004). Even more advanced treatments (e.g. Hardy 1993; Hardy and Reynolds 2004), which illustrate some awareness of this issue, have tended not to describe solutions that are open to the secondary analyst such as the reader of published output.

We conclude that the quasi-variance calculations described by Firth offer an attractive solution to the reference category problem that can be operationalised by sociological researchers. This is because in standard software information from the variance covariance matrix of the parameter estimates can be extracted[27]. This information can then be plugged into Firth's web-based calculator and quasi-variances can be estimated.

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

[25] For the interested reader, a worked illustration of an accuracy comparison is shown on the examples files at our website www.longitudinal.stir.ac.uk/qv/ .
[26] The inaccuracy percentages which are reported by the QV calculator refer to the maximum amount of difference between the conventional and quasi-variance based calculation of the standard error of the difference between parameter estimates, ie the difference between the values calculated by equation (2) and equation (3) above.
[27] As we have noted above this is more straightforward in STATA than in SPSS.

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. In addition our excel calculator is available to assist sociologists when performing the calculations necessary to compare categories.

We hope that this paper 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 a wider sociological community. We are optimistic that researchers might see that this methodological development has clear advantages for drawing substantive inference. We are aware that it may take some time for the research community to ubiquitously adopt the practice of reporting quasi-variances alongside parameter estimates and standard errors. How we hope that this paper has begun to convince some sociologists of the benefits of this approach.

Bibliography

Allison, P. D. (1999) *Multiple Regression: A Primer*. London: Sage.

Berk, R. A. (2004) *Regression Analysis: A Constructive Critique*. London: Sage.

Blumer, H. (1956) 'Sociological Analysis and the 'Variable'', *American Sociological Review* 21(6): 683-690.

Chambers, M. L. and C.J. Skinner (eds) (2003) *Analysis of Survey Data*. New York: Wiley.

Connolly, P. (2006) 'The effects of social class and ethnicity on gender differences in GCSE attainment: a secondary analysis of the Youth Cohort Study of England and Wales 1997-2001', *British Educational Research Journal*. 32(1): 3-21.

Cramer, D. (2003) *Advanced Quantitative Data Analysis*. Maidenhead, Berks: Open University Press.

Dale, A. (2006) 'Quality Issues with Survey Research', *International Journal of Social Research Methodology* 9(2): 143-158.

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.

Fielding, J. L. and G.N. Gilbert (2006) *Understanding Social Statistics, 2nd Edition*. 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, Austria: R Foundation for Statistical Computing (http://www.warwick.ac.uk/go/qvcalc).

Firth, D. and R.X. de Menezes ‘Quasi-variances’, *Biometrika* 91(1): 65-80.

Gallie, D. (1988) *The Social Change and Economic Life Initiative: An overview.* Oxford: ESRC-SCELI Working Paper 1, Nuffield College.

Goldthorpe, J.H. (2000) *On Sociology: Numbers, Narratives, and the Integration of Research and Theory*. Oxford: Oxford University Press.

Halaby, C. N. (2004) ‘Panel models in sociological research: Theory into practice’, *Annual Review of Sociology*. 30: 507-544.

Hardy, M. (1993) *Regression with Dummy Variables*. London: Sage.

Hardy, M. and A. Bryman (eds) (2004) *Handbook of Data Analysis*. 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*. London: Sage.

Harkness, J., F.J.R. van de Vijver, and P.Ph. Mohler (eds) (2003) *Cross-Cultural Survey Methods.* New York: Wiley.

Harsl∅f. I. (2005) ‘'Integrative' or 'defensive' youth activation in nine European welfare states’, *Journal of Youth Studies* 8(4): 461-481.

Hedeker, D. (2005) ‘Generalized Linear Mixed Models’, in B. Everitt and D.C. Howell (eds.) *Encyclopaedia of Statistics in Behavioural Science*, New York: Wiley.

Jaccard, J. and R. Turrisi (2003) *Interaction Effects in Multiple Regression, 2nd Edition*, London: Sage.

Menard, S. (2001) Applied *Logistic Regression Analysis, 2nd Edition*. London: Sage.

Menezes, R.X. de (1999) *More Useful Standard Errors for Group and Factor Effects in Generalized Linear Models.* Oxford: PhD Dissertation, Department of Statistics, University of Oxford.

ONS - Office for National Statistics. Social Survey Division (2006), *General Household Survey, 2001-2002 [computer file]. 4th Edition*. Colchester, Essex: UK Data Archive [distributor], February 2006. SN: 4646.

Pahl, R. and D.J. Pevalin (2005) ‘Between family and friends: a longitudinal study of friendship choice’, *British Journal of Sociology* 56(3): 433-450.

R Development Core Team (2006) *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Ridout, M.S. (1989) 'Summarizing the results of fitting Generalized Linear Models to data fro designed experiments', in A. Decarli, B. Francis, R. Gilchrist and G. Seeber *Statistical Modelling: Proceedings of GLIM89 and the 4th International Workshop on Statisitical Modelling.* New York: Springer-Verlag.

Rose, D. and D.J. Pevalin (2003) *A Researcher's Guide to the National Statistics Socio-economic Classification.* London: Sage.

Sandu, D. (2005) 'Emerging transnational migration from Romanian villages', *Current Sociology* 53(4): 555-582.

Skrondal, A. and S. Rabe-Hesketh (2004) *Generalized Latent Variable Modelling: Multilevel, Longitudinal, and Structural Equation Models*. London: Chapman and Hall/CRC.

SPSS (2004) *SPSS for Windows, Release 13.0.1*. Chicago: SPSS Inc.

Stata (2005) *Intercooled Stata 9.0 for Windows.* College Station, TX: StataCorp LP.

Stolzenberg, R.M. and D.A. Relles (1997) 'Tools for intuition about sample selection bias and its correction', *American Sociological Review* 62: 494-507.

van de Werfhorst, H. G. (2005) 'Social background, credential inflation and educational strategies', *Acta Sociologica* 48(4): 321-340.

Widmer, E., J. Kellerhals and R. Levy (2004). 'Tyes of conjugal networks, conjugal conflict and conjugal quality', *European Sociological Review* 20(1): 63-77.