

Logistic Regression Models in Sociological Research

Vernon Gayle

University of Stirling, and
ISER, University of Essex
University of Stirling

Paul S. Lambert

20th April 2009 [Edition 1.1]

DAMES Node, Technical Paper 2009-1

Technical Papers of the DAMES Node: Data Management through e-Social Science, <http://www.dames.org.uk/publications.html>. DAMES is an ESRC Research Node, Ref: RES-149-25-1066, based at the Universities of Stirling and Glasgow (the National e-Science Centre, www.nesc.ac.uk). The DAMES Node is a component of the ESRC National Centre for e-Social Science (www.ncess.ac.uk).

Version history:

Edition 1.0 published on 31st March 2009

Edition 1.1 published on 20th April 2009

Acknowledgements

We thank Richard B. Davies for his extended advice and comments on the topics covered and on previous versions of this paper.



Contents

1. Introduction	3
<i>1.1 Empirical Example (Connolly, 2006)</i>	4
<i>1.2 Data Management</i>	4
<i>1.3 The Data</i>	5
2. Operationalising and Estimating Logistic Regression Models	8
<i>Standard errors</i>	8
<i>Logit and Probit</i>	9
<i>Parameterisation</i>	9
<i>Model Fitting Strategy</i>	16
<i>Interpreting the Effects of Categorical Explanatory Variables with Odds Ratios</i>	18
<i>Alternative Methods for Interpreting the Effects of Explanatory Variables</i>	24
3. Sample Enumeration Methods for Interpreting the Substantive Effects of Individual Explanatory Variables	25
4. Conclusion	33
References	36
Endnotes	40

1. Introduction

Many readers will be familiar with statistical modelling approaches to analysing social survey data. Statistical models offer sociologists an attractive method to summarize patterns in survey datasets (Dale and Davies, 1994; Goldthorpe, 2007). Sociologists have tended to employ regression models in order to explore the effects of multiple explanatory variables on an outcome of interest. Standard statistical modelling approaches are becoming increasingly widely known and in the UK sociology postgraduate students are routinely trained in these techniques.ⁱ Advances in software packages for statistical analysis (e.g. SPSS, 2008; Stata, 2007) and the rapid increases in the power of desktop computers have greatly expanded the possibilities for developing statistical models to analyse survey datasets. These advances have been coupled with increased accessibility to survey datasets, particularly through internet based links to national data archives.ⁱⁱ

Many sociological studies use logistic regression models to analyse data in which the outcome variable has two discrete categories, classically referred to as a 'binary' outcome. Examples of binary outcome measures are legion, and measures frequently either take the form of no or yes, or relate to the presence or absence of some condition.ⁱⁱⁱ Within sociology there is a preference for logistic regression models when data with a binary outcome are modelled but, by contrast, the preference within economics is for the probit model. The logit and probit models are usefully considered as two specific models in the wider class of models termed 'Generalised Linear Mixed Models' (GLMM) (see Hedeker, 2005). Nevertheless Liao (1994) argues that given the similarities between the logit and probit models, either model will lead to identical substantive conclusions in most applications^{iv}. Because of its ubiquity within sociology and because of some specific problems associated with interpreting logistic regression models we concentrate on these models in the present paper. We use the terminology 'logistic regression' and 'logit models' interchangeably in this paper.

Even a cursory examination of the statistical models presented in sociological research in the UK would lead to the conclusion that many sociologists end their interpretation of modelling results at reporting signs (directions of effects) and statistical significance. In this paper we argue that there are benefits for analyses if effort is put into a fuller understanding of what is often referred to as the 'right hand side' of the equation (the estimated coefficients of the explanatory variables). We argue that placing more effort into interpreting modelling output will ultimately lead to better understanding of what Goldthorpe (2007) terms as 'empirical regularities' in the social world.

In a recent article two of the authors advanced an emerging statistical technique to assist the presentation and interpretation of statistical models that include categorical explanatory variables (Gayle and Lambert, 2007). This present paper dovetails with the previous paper and it focuses on some of the problems associated with modelling binary outcomes and presenting and interpreting their results in sociological studies. It is aimed at sociologists who analyse data with binary outcomes with statistical models, and those who read work which employs these techniques.

1.1 Empirical Example (Connolly, 2006)

We use an example from a recent empirical paper, Connolly (2006) which models the effects of gender, ethnicity and social class on General Certificate of Secondary Education (GCSE) attainment. The paper explored data from three Cohorts (9, 10 and 11) of the Youth Cohort Study of England and Wales (YCS). Three logistic regression models were fitted and presented in the paper (one to each cohort of data).^v The outcome measure is a discrete binary measure; whether or not the young person obtained five or more GCSEs at grades A*- C at the end of Year 11 (when they reached the end of compulsory education). The three explanatory measures included in the logistic regression models are measures of gender, ethnicity and social class.

We have chosen this example for a number of reasons. First, standard logistic regression models are applied to large-scale survey data. Second, the YCS is publicly available to academics and therefore researchers are in the position to replicate the original work and the work that we will develop below. Third, the outcome measure, five or more GCSEs at grades A*-C is routinely used in educational research and is also a recognised benchmark in official education statistics.^{vi} Fourth, the three explanatory variables are ‘key’ variables, by which we mean they are both important measures, and they are routinely included within a wide range of sociological analyses. Finally, we consider that whilst Connolly (2006) conducts a thoughtful analysis (of the effects of gender, ethnicity and social class on GCSE attainment), the development, presentation and interpretation of the models illustrate some potential problems with employing statistical models to binary outcome data. The focus of this paper is to provide some comments and solutions to these problems.

1.2 Data Management

Despite the increased availability of survey datasets, replicating analyses remains a major obstacle. In social research, the basic problem is that subsequent researchers frequently struggle to replicate models because there is little or no information available to them on how the data has been prepared to enable (or facilitate) statistical models being fitted. Nevertheless Freese (2007), amongst others, has argued that there are few compelling reasons why resources for the replication of sociological research results should be less widely delivered than comparable resources in other disciplines, where the tradition of documentation for the purposes of replication is well established.

Dale (2006) outlines a persuasive protocol for communicating data quality issues, and supporting replication, to readers of research and we hope that these ideas will be taken up more widely. With the now widespread availability of access to the internet researchers should be able to make such information available to others. One obvious strategy is to make available software syntax files (e.g. a .do file in Stata). We have deposited a Stata .do files at www.dames.org.uk/surveys/y/cs/ that perform the data preparation and analysis presented in this paper.^{vii}

Experienced survey data analysts are familiar with the importance of good practices such as keeping clear audit trails when analysing survey data. Using syntax files to process commands in statistical software packages makes a clear contribution to this

aim. A clear audit trail is essential to the replication of data by other analysts, but also to the replication of results for their modification by the primary analysts (for example attending to changes suggested by referees and examiners).

In the ongoing ESRC funded DAMES Node^{viii} two of the authors have surveyed a wide range of survey data training resources. We have reached the conclusion that at the current time the emphasis is firmly directed toward analysing data. By contrast experienced data analysts are acutely aware that statistical analyses (e.g. model building) come after, usually very lengthy, ‘data management’ activities.

We employ the term ‘data management’ to refer to tasks which generally involve ‘preparing’ survey data for social science analysis. Typically, this involves the sociologist performing tasks like recoding individual variables, but it can also involve more complex reorganisations of datasets (e.g. matching and merging files).^{ix} In our combined experience these tasks are usually much more time consuming than initially anticipated. For instance, the Stata command file “Connolly_replication.do”, available on our website, illustrates the length of the ‘data management’ process associated with even a simple activity like preparing a real survey dataset to fit a logistic regression model with only a small number of explanatory variables. Such complex preparation requirements are, moreover, often sufficient to dissuade applied researchers from engaging with suitable data resources or from taking appropriate steps in data preparation.

Although the topics of data management and preparation, and its documentation for the purposes of replication, have historically been neglected from methodological publications in social survey research, there is increasing contemporary interest in the process. We are heartened to be able to direct the reader to several recent texts aimed at mainstream audiences which confront these issues in a more complete manner. Kohler and Kreuter (2009), Treiman (2009) and Levesque (2008) offer data analysis texts which place emphasis on conducting and recording data preparation tasks. Long (2009) gives a highly focused account of the data analysis process itself.

1.3 The Data

The Youth Cohort Study of England and Wales (YCS) is an important data source. It is a large-scale nationally representative survey of young people that began in the mid 1980s. The survey tracks a representative cohort of young people when they reach the minimum school leaving age. The YCS is a panel study and usually three sweeps of data are collected on each cohort. The study collects detailed data on qualifications and educational experiences, employment and training. Information is also collected on the young person’s personal circumstances, their family and home life, and to a limited extent their aspirations and attitudes.

Our initial intention was to work with YCS Cohort 11 data because it is more recent (this cohort sat GCSE exams in Year 11 in 2001). We have been unable to replicate the results for YCS Cohort 11 presented in Connolly (2006) however. Therefore we have concentrated on developing analyses of YCS Cohort 9 because we have been able to replicate these results. Respondents in YCS Cohort 9 completed compulsory education (Year 11) in 1997. The dataset is available from the UK Data Archive.^x

General Certificates of Secondary Education (GCSE) were introduced in the late 1980s (Department of Education 1985; Mobley *et al.* 1986; North 1987).^{xi} They are the standard qualification for pupils in England and Wales. They are usually a mixture of assessed coursework and examinations. Generally each subject is assessed separately and a subject specific GCSE awarded. It is usual for pupils to study for about nine subjects, which will include core subjects (e.g. English, Maths and Science) and non-core subjects. For example the mean number of GCSE exam courses studied for respondents in YCS9 was eight.^{xii} GCSEs are graded in discrete ordered categories. The highest being A*, followed by grades A through to G.

Table 1 reports some descriptive characteristics of the YCS Cohort 9 sample. The outcome is a discrete binary measure, taking the value 1 if the respondent attained five or more GCSE at grades A*- C at the end of Year 11 (when they reached the end of compulsory education) and taking the value 0 otherwise. The national result for 1997 was 45.1% (DfCSF, 2009, Table 1).

Three key explanatory measures are included in the logistic regression models: gender, ethnicity and social class. The ethnicity measure broadly follows the 1991 UK Census categories (see Bulmer, 1996). The social class variable is a measure based on parental occupations. This is a derived measure that was deposited with the dataset. This measure is similar, but not identical, to the Registrar General's Classification (see Leete and Fox 1977). Our overall position is that sociologists should use published and recognised occupation based classifications. This is because they can be replicated and greatly enhance the possibility of comparisons between studies (Lambert *et al.*, 2007). Lambert and Bihagen (2007) provide a recent review of a wider range of alternative measures. Gayle, Lambert and Murray (2009) illustrate analyses of multiple cohorts of YCS data with a range of different established classifications based on parents' occupations.

The outcome variable and the three explanatory variables have been coded and organised to enable replicating the model of YCS Cohort 9 presented in Connolly (2006). The Stata .do file "ygs_data_1.do" available on our website can be used to organise the data.

Table 1 Characteristics of the YCS 9 Sample

Outcome Variable¹	%	
Less than 5 GCSE Passes	53.8	
5+ GCSE Passes at grades A*-C	46.2	
Explanatory Variables²	Weighted <i>n</i>	Proportion Attaining 5+ GCSEs
<i>Gender</i>		
Female	7269	0.51
Male	7393	0.42
<i>Ethnicity</i>		
Chinese	74	0.67
Indian	437	0.53
White	12894	0.47
Bangladeshi	122	0.33
Pakistani	312	0.29
Black	297	0.29
<i>Parental Social Class</i>		
Professional / Managerial	3049	0.69
Other Non-Manual	2830	0.60
Skilled Manual	4698	0.40
Semi-Skilled Manual	1702	0.32
Unskilled Manual	610	0.20

1. The weighted percentage concurs with the outcome variable reported in Connolly (2006), Table 2, p.10.

2. These numbers (weighted *n*) concur with the explanatory variables reported in Connolly (2006), Table 1, p.7.

2. Operationalising and Estimating Logistic Regression Models

Table 2 reports the results of two logistic regression models (Model 1 and Model 2). The outcome modelled is whether or not the young person (the YCS survey respondent) attained at least five GCSEs, at grades A* to C, in year 11 (at the end of compulsory school). The model includes three explanatory variables, gender, ethnicity and social class. In standard regression models, interpretation problems can arise when explanatory variables are highly correlated ('colinear'), but in the current example the three explanatory variables used have weak overall associations and satisfy standard tests for multicollinearity.^{xiii} Accordingly, we would conventionally expect the two logistic regressions to generate stable coefficient estimates. In the following section we comment on the features of the Models shown in Table 2 in the context of the calculation of coefficient standard errors; the comparison between logit and probit models; and the parameterisation of explanatory variables involved.

Standard errors

Columns 1 and 2 of Table 2 report the parameter estimates (β) and the standard errors of a logistic regression model reported in Connolly (2006) (Model 1). These results have been replicated using SPSS (version 16). This is a standard logistic regression model fitted to weight survey data. Model 2 is reported in columns 4 and 5 of Table 2. This model is a logistic regression model fitted in the survey data analysis suite (svy) within Stata (version 10).

Our first observation is that whilst the parameter estimates reported in columns 1 and 4 are identical the standard errors reported in columns 2 and 5 are not. This is because a standard logit model has been fitted to weighted data in SPSS, and this software does not correctly estimate the standard errors for models estimated on weighted data. In some instances this may be overlooked, however the calculation of statistical tests, most commonly p values and the construction of confidence intervals, can be adversely effected. The survey data analysis suite (svy) within Stata is specifically designed to analyse survey data sets and therefore appropriate (linearized) standard errors are computed and reported in column 5 of Table 2.^{xiv} In this present example the parameter estimates for young people of Bangladeshi and Pakistani origins in Model 1 are reported as being more significant (Table 2 column 3) than in Model 2 (Table 2 column 6). This is because the correct standard errors associated with these parameters have not been computed in SPSS.

The use of sampling weights in survey data analysis are often contentious. Whilst prescriptions differ, it is clearly preferable to employ a software package that provides correct standard errors based on a suitable computational method.^{xv} This is likely to become increasingly important as datasets with complex survey designs become increasingly available.^{xvi} In our experience, Stata is a particularly suitable package for the analysis of complex social survey data which may include the use of sampling weights.^{xvii}

Logit and Probit

We have asserted that the logit and probit model usually lead to the same substantive inference, because they are largely mathematically equivalent. Amemiya (1981) proposes a simple transformation of estimates between logit and probit models of 1.6. $\beta_{\text{logit}} = (\beta_{\text{probit}} * 1.6)$ and $\beta_{\text{probit}} = (\beta_{\text{logit}} / 1.6)$. Alternatively, Aldrich and Nelson (1984) suggest a scaling factor of $\pi / \sqrt{3} = 1.814$. Liao (1994) asserts that the most accurate value of the factor lies somewhere in the neighbourhood of these two values. If model 2 is re-estimated as a probit model the coefficient for females is estimated as $\beta_{\text{probit}} = 0.248$. The difference between the estimate of *Female* β_{probit} (0.248) and *Female* β_{logit} (0.405) is 1.629 in this particular instance. As in this instance, in our experience it is more generally the case that the substantive conclusions drawn from logit and probit models of the same outcome will generally be identical. Nevertheless Liao (1994) suggests that there are cases when probit and logit results differ substantially, for example when there are an extremely large number of observations heavily concentrated in the tails of the distribution. Here we would advise sociologists to take extra care in choosing the suitable model after comparing results from both.

Parameterisation

We argue that Model 1 and Model 2 (Table 1) are sub-optimally parameterised in how the categorical explanatory variables have been operationalised. We suspect that the particular parameterisation has been chosen because it is the default offered by SPSS. The software by default chooses the last category of a multiple category explanatory variable as the reference (or base) category. To help illustrate this issue Table 3 reports the results of Model 2 with quasi-variance (QV) standard errors and 95% comparison intervals.^{xviii}

In Figure 1 we plot the parameter estimates for each category of the ethnicity variable using the conventional 95% confidence intervals (hollow circles) taken from Model 1. None of the other ethnic categories have 95% confidence intervals that overlap with zero, the estimate for the base category (i.e. Black respondents). Therefore a conventional interpretation is that all of the other ethnic groups are considered to be significantly different (to Black young people). Whilst this interpretation is appropriate, in a narrow sense, it is beguiling, since it tells us little about the overall effects of ethnicity in the sample population.

The error arises because using the Black category as the reference category is problematic due to its small size (in the model there are 12789 cases but only 195 Black respondents). The effect of being Black is imprecisely estimated by the model, however this is not immediately apparent because it is the reference category and conventionally a standard error is not estimated.^{xix} The plot of the parameter estimates with 95% quasi-variance based comparison intervals (hollow diamonds) illustrate this more clearly. In this plot the Black category is seen to have a relatively wide comparison interval and we therefore we argue that methodologically it is sub-optimal as the reference category in this model.

We also consider that this category is a poor choice of reference category substantively. This is because respondents who identify themselves as either African,

Caribbean or Other Black are combined within this category.^{xx} It is recognised in ethnicity research that pupils from Black African and Black Caribbean backgrounds have quite different patterns of educational attainment (see Gillborn and Gripps 1996). This is acknowledged by Connolly (2006, p.8), whose choice to combine these groups is undertaken to avoid having small sample sizes in some categories of the explanatory variable.^{xxi} We strongly recommend that analysts should think hard about which category is used as a reference category and should be particularly cautious when combining categories in analyses because this can have substantial consequences for subsequent substantive interpretations.

We argue that the computation of quasi-variance based standard errors, and plotting 95% comparison intervals is attractive when communicating the effects of categorical explanatory variables which are common in sociological research. In this example the plot in Figure 1 is particularly useful as it provides a means by which a reader can gain a clearer understanding of the overall effects of ethnicity and can readily compare ethnic groups (even when one group is not the reference category). For instance, the quasi-variance plot in Figure 1 shows that Chinese and Indian young people have increased log odds of attaining 5+ GCSEs compared with White young people, and that Chinese young people have significantly higher log odds than their Indian counterparts, but this could not be seen by simply plotting conventional 95% confidence intervals.

In **Figure 2** we plot the parameter estimates for each category of the social class variable and conventional 95% confidence intervals (hollow circles). The plot suggests that the unskilled manual social class group is sub-optimal as a reference category because this category is relatively small ($n=590$). The structural order of the categories is interrupted in this parameterisation and therefore the monotonically decreasing (or step down) substantive effect of being from a less advantaged social class background is occluded.

The parameterisation of Model 3 (Table 4) is more appropriate and improves upon the parameterisation of Model 2. In this example more substantively and methodologically appropriate reference categories have been chosen. When plotted in Figure 3, it can be seen that with a better parameterised model the overall ethnicity effect is illustrated more clearly. In such instances, plotting quasi-variance comparison intervals is an effective means of presenting results (it is additionally beneficial to perform formal tests on the difference between pairs of coefficients, which can achieve the same effect). We return to Model 3 later.

Table 2 Logistic Regression Models: 5+ GCSEs (A*-C) Year 11 YCS Cohort 9

	1 Model 1 Logit model Connolly (2006)			4 Model 2 Stata survey (svy) logit model		
	B	Standard Error	Probability	B	Linearized Standard Error	Probability
<i>Gender</i>						
Female	0.405	0.038	<0.001	0.405	0.039	<0.001
Male	0.000			0.000		
<i>Ethnicity</i>						
Chinese	2.002	0.341	<0.001	2.002	0.377	<0.001
Indian	1.066	0.193	<0.001	1.066	0.208	<0.001
White	0.643	0.159	<0.001	0.643	0.171	<0.001
Bangladeshi	0.766	0.332	0.021	0.766	0.345	0.026
Pakistani	0.531	0.230	0.021	0.531	0.245	0.030
Black	0.000			0.000		-
<i>Social Class</i>						
Professional / Managerial	2.192	0.110	<0.001	2.192	0.109	<0.001
Other Non-Manual	1.773	0.110	<0.001	1.773	0.108	<0.001
Skilled Manual	0.932	0.107	<0.001	0.932	0.104	<0.001
Semi-Skilled Manual	0.576	0.115	<0.001	0.576	0.113	<0.001
Unskilled Manual	0.000	-	-	0.000		-
Constant	-2.208	0.189		-2.208	0.198	

Note: Estimates are reported to three decimal places to provide consistency with Connolly (2006).

Table 3 Model 2 Survey Logistic Regression Model (Stata): 5+ GCSEs (A*-C) Year 11 YCS Cohort 9

	1	2	3	4	5	
	B	Linearized standard error	Probability	QV based standard error	QV based 95% comparison intervals	
					Lower	Upper
<i>Gender</i>						
Female	0.405	0.039	<0.001	-	-	-
Male	0.000	-		-	-	-
<i>Ethnicity</i>						
Chinese	2.002	0.377	<0.001	0.3364	1.343	2.662
Indian	1.066	0.208	<0.001	0.1201	0.830	1.301
White	0.643	0.171	<0.001	0.0203	0.603	0.683
Bangladeshi	0.766	0.345	0.026	0.2998	0.179	1.354
Pakistani	0.531	0.245	0.030	0.1761	0.186	0.877
Black	0.000	-	-	0.1701	-0.333	0.333
<i>Social Class</i>						
Professional /						
Managerial	2.192	0.109	<0.001	0.044	2.107	2.278
Other Non-Manual	1.773	0.108	<0.001	0.042	1.691	1.854
Skilled Manual	0.932	0.104	<0.001	0.031	0.871	0.993
Semi-Skilled Manual	0.576	0.113	<0.001	0.053	0.472	0.680
Unskilled Manual	0.000	-	-	0.100	-0.195	0.195
Constant	-2.208	0.198	-	-	-	-

n=12789; Non-weighted logistic regression model, Deviance =15993, Pseudo R²=0.070.

Figure 1 Ethnicity Effects: Model 2 Survey Logistic Regression Model (Stata)

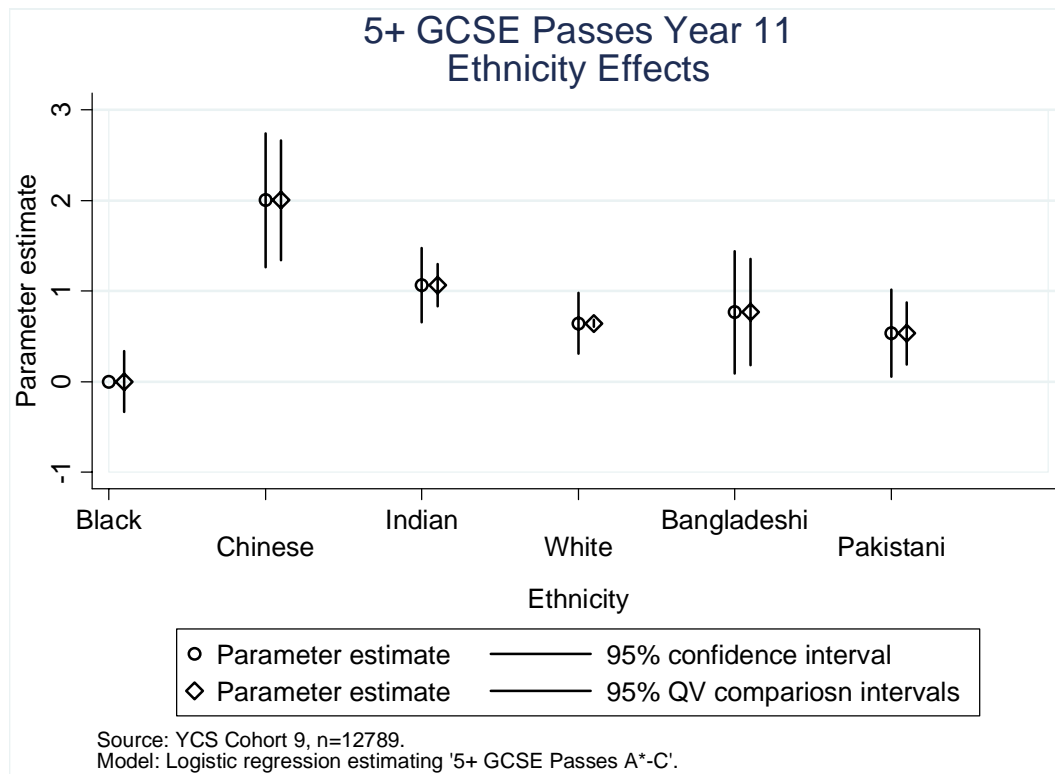


Figure 2 Social Class Effect: Model 2 Survey Logistic Regression Model (Stata)

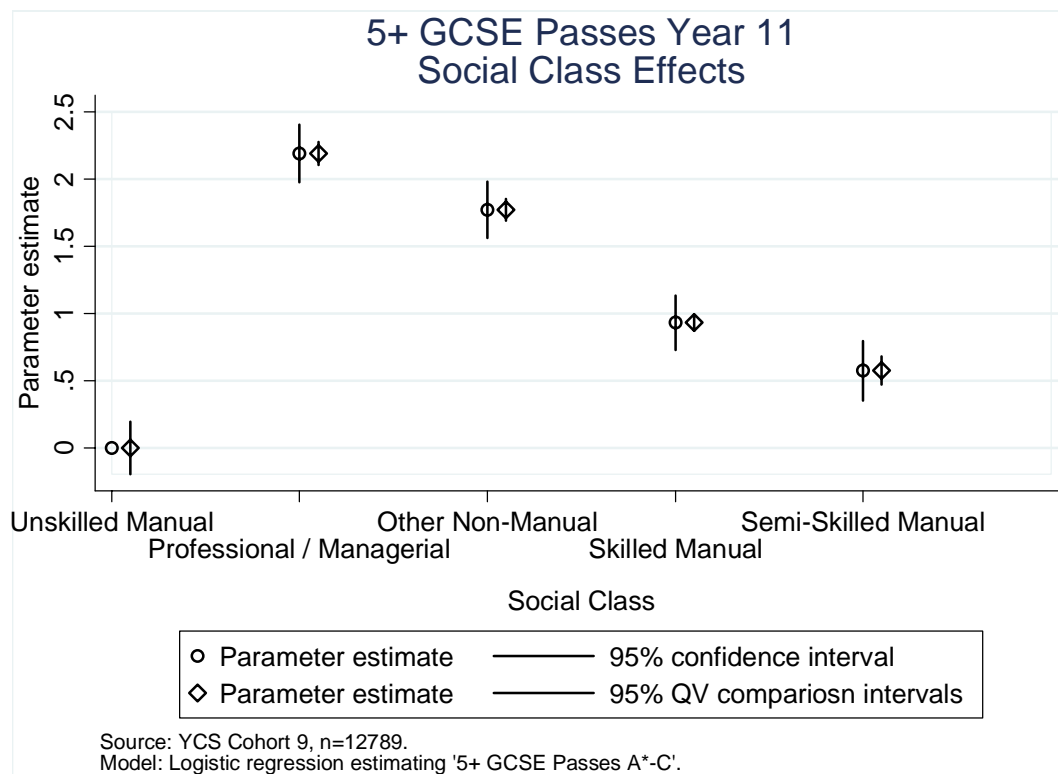
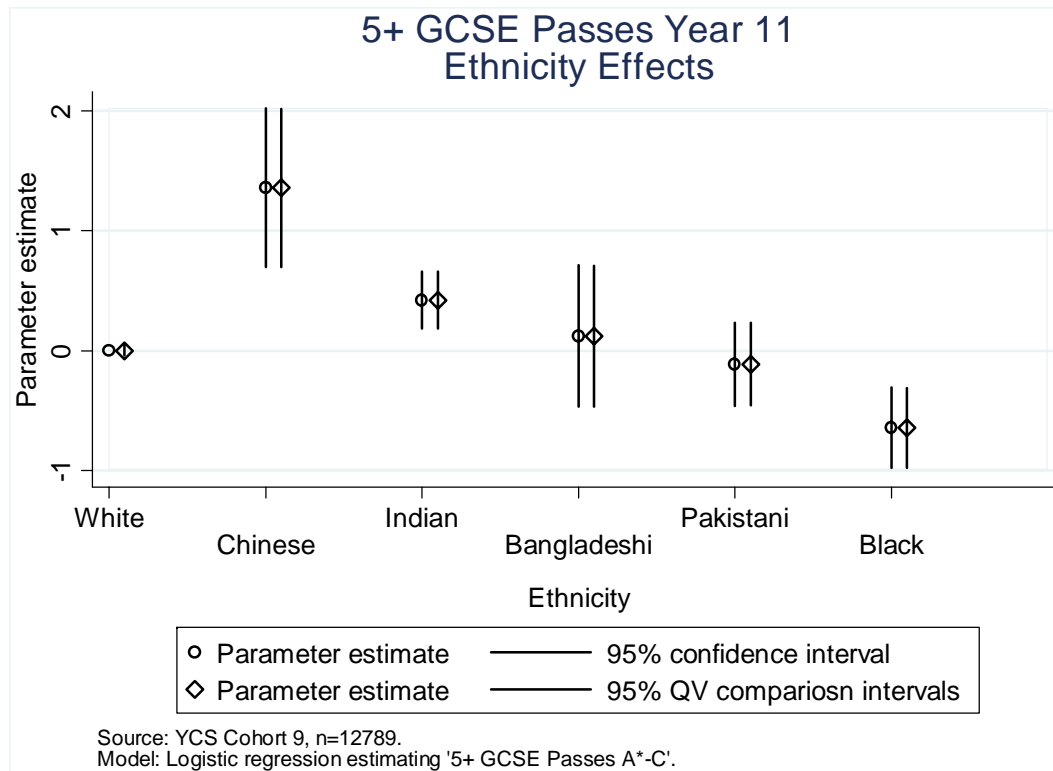


Table 4 Model 3 Survey Logistic Regression Model (Stata): 5+ GCSEs (A*-C) Year 11 YCS Cohort 9

	B	Linearized s.e.	p	QV based s.e.	95% comparison intervals	
					Lower	Upper
Gender						
Female	0.000	-	-	-	-	-
Male	-0.405	0.039	0.000	-	-	-
Ethnicity						
Chinese	1.359	0.337	0.000	0.336	0.700	2.019
Indian	0.423	0.122	0.001	0.120	0.187	0.658
White	0.000			0.020	-0.040	0.040
Bangladeshi	0.123	0.301	0.682	0.300	-0.465	0.711
Pakistani	-0.112	0.177	0.529	0.176	-0.457	0.233
Black	-0.643	0.171	0.000	0.170	-0.977	-0.310
Social Class						
Professional / Managerial	0.000	-	-	0.044	-0.085	0.085
Other Non-Manual	-0.420	0.060	0.000	0.042	-0.502	-0.338
Skilled Manual	-1.260	0.053	0.000	0.031	-1.321	-1.199
Semi-Skilled Manual	-1.616	0.069	0.000	0.053	-1.720	-1.512
Unskilled Manual	-2.192	0.109	0.000	0.100	-2.387	-1.997
Constant	1.032	0.048	-	-	-	-

n=12789; Non-weighted logistic regression model, Deviance =15993, Pseudo R²=0.070.

Figure 3 Ethnicity Effects: Model 3 Survey Logistic Regression Model (Stata)



Model Fitting Strategy

We report the series of models that could potentially lead to the adoption of Model 2 in Table 5. This table illustrates a typical ‘model building’ process in a social science analysis, and indicates a particular ‘model building strategy’, namely, the incremental expansion of the number of explanatory variables.

In our view the development of a clear and well thought out model fitting strategy is too often overlooked in sociological analyses. In models with a large number of explanatory variables, consideration of the model fitting strategy is especially important. We contend that ideally the model building process should, at all stages, be theoretically guided, but we are also aware that much statistical modelling in sociology is of an exploratory nature.^{xxii} Some software packages provide automated methods for model building, for example stepwise regression. We would seldom advocate such methods as they are guided by statistical rules which can easily generate a model of poor substantive quality.

We highlight that the order in which explanatory variables are added and which variables are omitted requires thought. It is common for coefficient effects to change substantially when other relevant explanatory variables are added. We generally advocate building regression models slowly, by adding and subtracting explanatory variables. This usually provides the data analyst with better insights into the right hand side of the equation. For instance, whether coefficient effects increase, decrease or are unaltered by other explanatory variables can itself be of substantive interest.

Another common model building strategy involves entering all available explanatory variables in a first step, then subsequently removing selected explanatory variables. This is sometimes referred to as ‘stewing pot’ method, because all of the ingredients are thrown in together. This approach can give useful insight into the net effects of variables in the context of many other explanatory factors. However a significant danger of such an approach is that it can occlude simpler underlying structures within the model, and may sometime prioritise certain coefficient effects for arbitrary reasons (such as of functional form, or due to model ‘overfitting’). Accordingly, we argue that this model building strategy should ordinarily be avoided, or at the very least treated with caution.

Which statistical model is ultimately favoured (such as the ‘full model’ in Table 5) is also an issue of model building strategy. The choice is not a trivial matter. Despite the advances in the speed of computers and improvements in software, model choice is not a mechanical outcome of the model fitting process, even for relatively simple statistical models. Our advice is that analysts should exploit and compare several models fitting strategies before adopting a final model.^{xxiii} Although space requirements can prevent the presentation of the results of intermediate models in final reports, for the purposes of replication they could and should be made available via the internet.

Table 5 Model Building: Parameter Estimates Model 2 Survey Logistic Regression Model (Stata): 5+ GCSEs (A*-C) Year 11 YCS Cohort 9

	Null Model	Gender	Ethnicity	Class	Gender & Ethnicity	Gender & Class	Ethnicity & Class	Full Model
<i>Gender</i>								
Female		0.351			0.334	0.415		0.405
Male		0.000			0.000			0
<i>Ethnicity</i>								
Chinese			1.600		1.620		1.950	2.002
Indian			1.042		1.041		1.060	1.066
White			0.787		0.786		0.638	0.643
Bangladeshi			0.190		0.202		0.761	0.766
Pakistani			0.004		0.017		0.525	0.531
Black			0.000		0.000		0.000	0.000
<i>Social Class</i>								
Professional /Managerial				2.188		2.206	2.178	2.192
Other Non-Manual				1.766		1.783	1.759	1.773
Skilled Manual				0.961		0.959	0.936	0.932
Semi-Skilled Manual				0.602		0.600	0.580	0.576
Unskilled Manual				0.000		0.000	0.000	0.000
Constant	-0.153	-0.328	-0.908	-1.367	-1.075	-1.581	-1.996	-2.208
Deviance	20004	19938	19199	16506	19141	16428	16066	15993
Pseudo R ²	0.00	0.00	0.01	0.06	0.01	0.07	0.07	0.07

Interpreting the Effects of Categorical Explanatory Variables with Odds Ratios

In our experience the effects of the individual explanatory variables in logistic regression models are rarely well described in sociological studies (and habitually mystify sociology students). The problem can be best understood in comparison with linear regression models. In the case of a linear regression model the effect of a continuous explanatory variable x_1 can be interpreted in a relatively straightforward manner. A one unit change in x_1 leads to a change in the y variable equal to the value of β_1 .

There is no equivalent simple interpretation of the effect of a single explanatory variable in the logistic modelling framework. This is because the outcome variable, which takes either the value 0 or 1, is transformed to map onto the range of probabilities. The result is that the β_1 is the effect that a change in x_1 has on the log odds of the outcome variable y taking the value 1.^{xxiv} Moreover, for categorical explanatory variables the β associated with category effects should be thought of as the effect on the log odds of moving from the reference (or base) category to the particular category (or level) of the X variable.

In logistic regression models, the log odds scale is not readily interpretable and communicating the effects of a single explanatory variable is much less straightforward than in linear regression models. The use of odds ratios are frequently advocated in methodological text books. Our overall position is that odds ratios should be avoided when interpreting the effect of explanatory variables and when communicating results from logistic regression models. This point may at first appear extreme but we will illuminate it further through the example from Connolly (2006).

The calculation of odds and odds ratios is relatively straightforward whereby the change in odds associated with a coefficient equates to the exponential of its β value. Most mainstream statistical software packages will report this result alongside other measures. Turning our attention to the effects of gender in Model 2, Table 6 column 1 reports that for females $\beta=0.405$ and correspondingly odds $=\exp(0.405)=1.499$ (column 2). This figure represents the increased odds that female respondents have of attaining five or more GCSEs. In contrast, turning to the effects of gender in Model 3, Table 6 column 3 reports that for males $\beta=-0.405$ and correspondingly the odds $=\exp(-0.405)=0.667$ (column 4). This figure represents the decreased odds that male respondents have of attaining five or more GCSEs. There is an obvious symmetry between these results, the odds ratio for females compared with males $=1/1.499=0.667$, and males compared with females $1/0.667=1.499$. Indeed, whenever the explanatory variable is binary the results are symmetrical, therefore which category of the explanatory variable is coded either as 0 or 1 is largely unimportant.

So far the use of odds and odds ratios may seem innocuous. However what do these figures actually mean? For the moment we will turn our attention from the effect of gender to the effects of ethnicity (Table 6 column 1). Connolly (2006) reports that 'Chinese respondents were found to be about seven times more likely to gain five or more GCSEs than Black respondents'. This is a common interpretation of the effects of an explanatory variable in logistic regression models, but we will demonstrate that it is misleading. We will begin by stating that, at a simple level, this interpretation does not obviously chime with the observed data. The (weighted) observed overall

percentage of young people attaining 5+ GCSEs is 46%. Twenty nine percent of Black young people attained 5+ GCSEs compared with 67% of Chinese young people (see Table 1).

The calculation of predicted probabilities for a logistic regression model with three explanatory variables is expressed in equation 1

$$\hat{p}_i = \frac{\exp(\beta_0 + \beta_1 + \beta_2 + \beta_3)}{1 + \exp(\beta_0 + \beta_1 + \beta_2 + \beta_3)} \quad (1)$$

In the current example the model includes the constant (β_0), gender (β_1), ethnicity (β_2) and social class (β_3).

Therefore the predicted probability of a male, Black respondent from an unskilled manual social class background attaining 5+ GCSEs is .10 .

$$\hat{p}_i = \exp(-2.208 + 0 + 0 + 0) / (1 + \exp(-2.208 + 0 + 0 + 0))$$

$$\hat{p}_i = .10$$

And the predicted probability of a male, Chinese respondent from an unskilled manual social class background attaining 5+ GCSEs is .45 .

$$\hat{p}_i = \exp(-2.208 + 0 + 2.002 + 0) / (1 + \exp(-2.208 + 0 + 2.002 + 0))$$

$$\hat{p}_i = .45$$

Table 7 Predicted Probabilities of Attaining 5+ GCSEs (A*-C) Year 11 by Gender, Ethnicity and Social Class reports the predicted probabilities for all combinations of the three explanatory variables in Model 2. Yet if the probability of a male, Black respondent from an unskilled manual social class background attaining 5+ GCSEs is .10, we might reasonably expect that the predicted probability for a comparable Chinese respondent should be much higher than .45 (since the reported effect of Chinese ethnicity was about seven times more likely to gain five or more GCSEs than a Black respondent).

One problem that bedevils the interpretation of odds as described above is easily illustrated. Table 8 reports the conversion of log odds to odds and to probabilities. Consider 0 on the log odds scale, which can be converted to an odds of 1 (i.e. $\exp(0)$), and transformed into a probability ($p=.5$) through the formulae $\text{odds} / (1 + \text{odds})$. Let us examine the relationship between log odds (Table 8 column 2) and their corresponding probabilities (column 3). Log odds from 0 to 4.60 correspond to probabilities between 0.50 and 0.99. Log odds from 0 to -4.60 correspond to probabilities between 0.50 and 0.01. Now let us compare the relationship between probabilities (Table 8 column 3) with odds (column 1). Probabilities ranging from 0.01 to 0.50 are compressed on the odds scale (column 1) from 0.001 to 1.00 whereas probabilities from 0.50 to 0.99 range from 1.00 to 99 on the odds scale. The relationship is not linear, and positive estimates on the odds scale are not constrained in the same manner as negative estimates.

Therefore we strongly warn against plotting odds ratios because effects are distorted, positive effects being stretched and negative effects compressed. The lack of a linear

relationship between odds and probability, in our view, lays down a banana skin that can trip the unwitting data analyst when communicating the results of logistic regression models. In our experience this issue is exacerbated in poorly parameterised logistic regression models when the contrast between a particular category and the reference category (for which the odds are artificially set to 1) may be imprecise due to a small sample size in these categories. For example in Model 2 the reference category Black young people contains 195 respondents and the Chinese category 61 respondents. The imprecision of the estimate for Chinese young people is reflected in the large standard error (0.377) (see Table 3 **Model 2 Survey Logistic Regression Model (Stata): 5+ GCSEs (A*-C) Year 11 YCS Cohort 9**). The estimate 2.002 lies within a rather wide 95% confidence interval

$$\text{c.i.} = 2.0002 \pm (1.96 * .377) = (1.263, 2.742)$$

This wide confidence interval is depicted in Figure 3. It is worthy of note that if the ‘true’ estimate was at the lowest point of this confidence interval, then the corresponding odds would be 3.535 which is a much more conservative estimate of the effect of being of Chinese origin. Conversely, if the ‘true’ estimate was at the highest point of this confidence interval, then the corresponding odds would be 15.517. As we have already suggested, an assertion along the lines of this group being ‘15 times more likely’ to attain 5+ GCSEs is clearly misleading. Because of the non-linear nature of the odds scale, large positive effects are misleadingly magnified. We consider that researchers should try to avoid reporting odds relating to the effects of single explanatory variables in logistic regression models, and exercise caution if they do report odds; this is particularly important when a category or level of an explanatory variable is imprecisely measured.

A more sensibly parameterised model can reduce this problem. For example in Model 3 (see Table 6 **Survey Logistic Regression Model 2 & Model 3: 5+ GCSEs (A*-C) Year 11 YCS Cohort 9**) White young people are the reference category with odds of 1. Young people of Indian origin have odds of 1.5 of attaining 5+ GCSEs. This is plausible in so far as the predicted probability for white young men from unskilled manual backgrounds is 0.17 and for their counterparts of Indian origin it is 0.24 (see Table 7 **Predicted Probabilities of Attaining 5+ GCSEs (A*-C) Year 11 by Gender, Ethnicity and Social Class**). Overall however the problem of the non-linearity of the odds scale bedevils the interpretation of positive effects in logistic regression models.

An insightful caution is issued by Drew, Gray and Sime (1992) in an earlier analysis of YCS data. Odds ratios are multiplicative and the positive effect of one explanatory variable can be reduced by the negative effect of another. This is illustrated in Model 3 (Table 6). Consider the positive effect of being of Indian origin (1.526). The odds of a young male of Indian origin and from a professional / managerial background attaining 5+ GCSEs is $2.805 * 1.000 * 1.526 * 1.000 = 4.280$. Whereas the odds of a young male of Indian origin and from the ‘other non-manual’ background attaining 5+ GCSEs is $2.805 * 1.000 * 1.526 * 0.657 = 2.812$. The positive effect of being of Indian origin is reduced by about 2/3 by the negative effect of being from the other non-manual background.

Overall we strongly advise that researchers should avoid using odds to interpret the effect of individual explanatory variables estimated by logistic regression models. If odds ratios must be used then terms like ‘higher’ and ‘lower’, and ‘increased’ and ‘decreased’, should be used to describe comparative effects. Researchers should avoid confusion by not conflating odds with probabilities, using terminology such as ‘more likely’ and ‘less likely’.

Table 6 Survey Logistic Regression Model 2 & Model 3: 5+ GCSEs (A*-C) Year 11 YCS Cohort 9

	1	2	3	4
	Model 2		Model 3	
	B	Odds	B	Odds
<i>Gender</i>				
Female	0.405	1.499	-0.405	0.667
Male	0.000	1.000	0.000	1.000
<i>Ethnicity</i>				
Chinese	2.002	7.404	1.359	3.893
Indian	1.066	2.904	0.423	1.526
White	0.643	1.902	0.000	1.000
Bangladeshi	0.766	2.151	0.123	1.131
Pakistani	0.531	1.701	-0.112	0.894
Black	0.000	1.000	-0.643	0.526
<i>Social Class</i>				
Professional / Managerial	2.192	8.953	0.000	1.000
Other Non-Manual	1.773	5.888	-0.420	0.657
Skilled Manual	0.932	2.540	-1.260	0.284
Semi-Skilled Manual	0.576	1.779	-1.616	0.199
Unskilled Manual	0.000	1.000	-2.192	0.112
Constant	-2.208	0.110	1.032	2.805

Table 7 Predicted Probabilities of Attaining 5+ GCSEs (A*-C) Year 11 by Gender, Ethnicity and Social Class
YCS Cohort 9 (Survey Logistic Regression Model 2)

	<i>Ethnicity</i>	<i>Social Class</i>				
		Professional / Managerial	Other Non-Manual	Skilled Manual	Semi-Skilled Manual	Unskilled Manual
<i>Gender</i>						
Males	Chinese	0.88	0.83	0.67	0.59	0.45
	Indian	0.74	0.65	0.45	0.36	0.24
	White	0.65	0.55	0.35	0.27	0.17
	Bangladeshi	0.68	0.58	0.38	0.30	0.16
	Pakistani	0.63	0.52	0.32	0.25	0.16
	Black	0.50	0.39	0.22	0.16	0.10
Females	Chinese	0.92	0.88	0.76	0.68	0.55
	Indian	0.81	0.74	0.55	0.46	0.32
	White	0.74	0.65	0.44	0.36	0.24
	Bangladeshi	0.76	0.68	0.47	0.39	0.22
	Pakistani	0.72	0.52	0.42	0.33	0.22
	Black	0.60	0.49	0.30	0.23	0.14

Table 8 Conversion of Log Odds, Odds and Probabilities

1	2	3
Odds	Log Odds (Logit Scale)	Probabilities
99.00	4.60	0.99
19.00	2.94	0.95
9.00	2.20	0.90
4.00	1.39	0.80
2.33	0.85	0.70
1.50	0.41	0.60
1.00	0.00	0.50
0.67	-0.41	0.40
0.43	-0.85	0.30
0.25	-1.39	0.20
0.11	-2.20	0.10
0.05	-2.94	0.05
0.01	-4.60	0.01

Alternative Methods for Interpreting the Effects of Explanatory Variables

One obvious solution to the obstacle of presenting odds to interpret the effects of individual explanatory variables is the presentation of probabilities, as we have shown in Table 7. In our experience it is often a very helpful way in which the data analyst can get to grips with the ‘action’ on the right hand side of the equation. We would however remind analysts that these probabilities are point estimates and appropriate caution should be exercised regarding their precision when substantive statements are being made.

Even in a fairly parsimonious model like Model 2 the three explanatory variables require a 2*6*5 cell table to show all possible combinations. In printed output (e.g. a journal article) it is seldom feasible to report a table for a model with a large number of categorical explanatory variables. As we suggest however, it would be possible for authors to publish such supporting information on the internet.

We have found it highly beneficial when communicating results from logistic regression models to non-technical audiences to provide a small number of illustrative examples using predicted probabilities (see Gayle, Berridge and Davies 2003). For example it might be useful to communicate that a white, female pupil from a professional / managerial background has a 74% chance of attaining 5+ GCSEs compared with a male pupil of a similar ethnic and social class background, whose chances are, on average, 9% lower. We have experienced that non-technical audiences appreciate this type of communication of results.

Presenting specific contrasts is useful but does not necessarily communicate the overall substantive effect of an individual explanatory variable. In our experience various suggestions surface from time to time and some are more satisfactory than others. One suggestion is that when communicating the results of a single explanatory variable the analysts set all of the values of the other explanatory variables to their modes and report the effects of changes in the value of the explanatory variable of interest. This can be an effective short cut, however it rests on the assumption that the modes of the other explanatory variables provide plausible groups for substantive comparisons.

A slight variation to this approach when the logistic regression model contains a large number of binary explanatory variables is to assess the impact of a change in a single explanatory variable whilst the other explanatory variables are each set to their means. This can be an effective short cut, however it rests on the assumption that the means of the explanatory variables provide suitable points of comparison. Some analysts might be disturbed by interpreting the effects of an explanatory variable in a formulation where, by analogy, the individual is 0.4 male and 0.6 female. Whilst this approach can be a useful heuristic device to get a better handle on the effects of an explanatory variable, in our experience when explaining the results of models to non-technical audiences this approach has the potential to raise concerns, if not lead to derision.

Gelman and Hill (2008) suggest dividing coefficients from logit models by 4 as a guide for assessing the effects of the β estimated for a given explanatory variable as a probability. They assert that $\beta/4$ provides a ‘rule of convenience’ for estimating the

upper bound of the predictive difference corresponding to a unit change in the explanatory variable. Returning to Model 2 the estimate for females $\beta_{\text{female}}=0.405$. This suggests a change of up to 10% ($0.405/4$) associated with being female. This is consistent with the predicted probabilities reported in Table 7. Gelman and Hill (2008) are careful to report that this is an approximation and that it performs best near the midpoint of the logistic curve (the range of effects). We believe that this has some merit as a rough and ready method of interpreting the effects of estimates and is a useful tool for many purposes.

3. Sample Enumeration Methods for Interpreting the Substantive Effects of Individual Explanatory Variables

In a methodological paper, Davies (1992), one of the authors illustrates the method of sample enumeration for interpreting the substantive effects of individual explanatory variables in non-linear models. The statistical proof is illustrated using the probit model framework. We have stated that logit and probit models should be considered as closely related alternative approaches for modelling binary outcomes, as both are from the wider GLMM family. Accordingly, this statistical theory readily extends to the logit model and this technique therefore has direct application for quantifying the substantive effects of individual explanatory variables in logistic regression models. Indeed Gayle, Berridge and Davies (2002) and Payne (1999) employ this technique to interpret results from standard logistic regression models. Davies, Elias and Penn (1992) and Davies (1994) use a version of the technique to interpret the results of logit panel models.

When models include a large number of explanatory variables quantifying and then communicating the effects of individual explanatory variables is especially problematic. Therefore, to illustrate the potential benefits of the sample enumeration approach we estimate a more comprehensive logistic regression model with additional explanatory variables. Table 9 reports the characteristics of the additional explanatory variables that are included in Model 4. We have chosen these variables because they are substantively plausible (see Gayle, Berridge and Davies, 2003) and they are not highly correlated with the three original explanatory variables (or each other) and they do not induce multicollinearity in the model. Model 4 is reported in Table 10.

Table 9 Characteristics of Additional Explanatory Variables YCS 9 Sample

Additional Explanatory Variables³	Weighted <i>n</i>	Proportion Attaining 5+ GCSEs
<i>Year 11 School Type</i>		
LEA Comprehensive	10497	0.41
Grant Maintained Comprehensive	2092	0.47
Selective School	496	0.94
Secondary Modern School	532	0.27
Independent School	1044	0.85
<i>Family Housing Tenure</i>		
Own Home	11255	0.53
Council Housing	2039	0.17
Housing Association	371	0.30
Private Landlord	467	0.34
<i>Parental Education</i>		
Non-Graduate Parent(s)	11479	0.40
Graduate Parent(s)	3183	0.69
<i>Family Composition</i>		
Non Lone Mother	11982	0.49
Lone Mother	2095	0.37
<i>Family Size</i>		
Less than 4 siblings in household	14315	0.47
4+ siblings in household	347	0.23

Table 10 Model 4 Survey Logistic Regression Model (Stata): 5+ GCSEs (A*-C) Year 11 YCS Cohort 9

	B	Linearized standard error	Probability	QV based standard error	QV based 95% comparison intervals	
					Lower	Upper
<i>Gender</i>						
Female	0.000					
Male	-0.498	0.042	<0.001	-	-	-
<i>Ethnicity</i>						
All Other Groups	0.000	-	-	0.010	-0.019	0.019
Chinese	1.361	0.449	0.002	0.010	1.341	1.380
Black	-0.439	0.196	0.025	0.004	-0.446	-0.431
<i>Social Class</i>						
Professional / Managerial	0.000	-	-	0.049	-0.095	0.095
Other Non-Manual	-0.166	0.065	0.011	0.045	-0.255	-0.077
Skilled Manual	-0.797	0.059	<0.001	0.034	-0.864	-0.730
Semi-Skilled Manual	-0.972	0.077	<0.001	0.058	-1.086	-0.858
Unskilled Manual	-1.413	0.119	<0.001	0.108	-1.625	-1.201
<i>Year 11 School Type</i>						
LEA Comprehensive Grant Maintained	0.000	-	-	0.0239	-0.047	0.047
Comprehensive	0.135	0.058	0.020	0.0526	0.032	0.238
Selective School	2.582	0.231	<0.001	0.2295	2.132	3.032
Secondary Modern School	-0.641	0.120	<0.001	0.1172	-0.871	-0.411
Independent School	1.586	0.129	<0.001	0.1269	1.337	1.835

Model 4 Survey Logistic Regression Model (Stata): 5+ GCSEs (A-C) Year 11 YCS Cohort 9 (continued)*

	B	Linearized standard error	Probability	QV based standard error	QV based 95% comparison intervals	
					Lower	Upper
<i>Family Housing Tenure</i>						
Own Home	0.000	-	-	0.028	-0.056	0.056
Council Housing	-1.114	0.076	<0.001	0.071	-1.253	-0.975
Housing Association	-0.445	0.150	0.003	0.147	-0.733	-0.157
Private Landlord	-0.449	0.144	0.002	0.141	-0.725	-0.173
<i>Parental Education</i>						
Non-Graduate Parent(s)	0.000	-	-	-	-	-
Graduate Parent(s)	0.617	0.056	<0.001	-	-	-
<i>Family Composition</i>						
Non Lone Mother	0.000	-	-	-	-	-
Lone Mother	-0.157	0.067	0.019	-	-	-
<i>Family Size</i>						
Less than 4 siblings in household	0.000	-	-	-	-	-
4+ siblings in household	-0.589	0.176	0.001	-	-	-
Constant	0.645	0.058	<0.001	-	-	-

n=12246; Non-weighted logistic regression model, Deviance =14104, Pseudo R²=0.138.

The principle behind the method of sample enumeration can be explained concisely. First the logistic regression model is estimated in a statistical software package in the usual fashion. In the current example Model 4 includes a constant (β_0), gender (β_1), ethnicity (β_2) and social class (β_3), year 11 school type (β_4), housing tenure, (β_5) parental education (β_6), family composition (β_7), family size (β_8).

An explanatory variable of substantive interest is then focused on. The technique facilitates the comparison between groups within (or levels of) this explanatory variable. A main group of interest is identified as a benchmark in comparisons. For example the benchmark group may be those young people from professional / managerial parental social class backgrounds, a group of young people who perform particularly well at GCSE (69% of this group attain at least 5+ GCSEs at grades A-C).

A comparison group which will be compared with the benchmark group (or category) is then extracted from the dataset (e.g. young people from other non-manual parental social class backgrounds). The estimated model is then applied to each individual in the extracted dataset with the variable of interest (social class) set to zero (i.e. $\beta_3=0$). Setting the explanatory variable of interest to zero is analogous to removing the particular effect, whilst recognising the effects of the other explanatory variables in the model.

$$\hat{p}_i(class) = \frac{\exp(\beta_0 + \beta_1 + \beta_2 + 0 + \beta_4 + \beta_5 + \beta_6 + \beta_7 + \beta_8)}{1 + \exp(\beta_0 + \beta_1 + \beta_2 + 0 + \beta_4 + \beta_5 + \beta_6 + \beta_7 + \beta_8)} \quad (2)$$

Summing the probabilities for each individual allows us to construct an estimated frequency (see Equation 2). In this example it is the number of young people that are estimated to attain 5+ GCSEs given their combination of other explanatory variables, other than the variable of interest (i.e. social class). From this estimated frequency we can easily compute an estimated proportion from the extracted sample. This we term as the ‘sample enumerated’ proportion. Using bootstrapping techniques it is possible to construct a pseudo-confidence interval around the expected frequency and therefore the sample enumeration proportion.

Row 1 of Table 11 reports the observed proportion of young people attaining 5+ GCSEs in each social class category. The overall proportion of young people attaining 5+ GCSEs is 0.49, it is 0.69 for those from professional / managerial backgrounds and 0.20 for those from unskilled manual backgrounds. The observed difference between those from unskilled manual backgrounds and their counterparts from professional / managerial backgrounds is 0.49 (Table 11 row 5).

The question that naturally follows, is how much of this observed difference (0.49) is due to the ‘direct’ effect of social class and how much of it is due to the other explanatory variables included in the model. An alternative, and more colloquial, way of expressing the same question is ‘what would happen if the class effect was removed?’ Or more specifically, what proportion of young people from unskilled manual backgrounds do we estimate can be expected to attain 5+ GCSEs if the class effect was set to zero, given the values of their other explanatory variables?

Returning again to Table 11, row 3 reports that ‘sample enumerated’ proportion for each group. We estimate a proportion of 0.46 for the unskilled manual class and we can conclude that this increase of 0.26 is due to the removal of the social class effect (row 6). Of the original observed difference in proportion for these two social class groups (0.49), 0.26 can reasonably be attributed to the ‘direct’ effect of social class and 0.23 to the effects of all other variables included in the model.

A further development on earlier uses of sample enumeration techniques is the estimation of confidence intervals around estimates. Using bootstrapping techniques it is possible to construct a pseudo-confidence interval around the sample enumeration proportion, which provides information on the precision of the estimate (Table 11 row 2 and row 4). The sample survey in the present example is large and the subsamples in each social class group are also relatively large. Therefore the sample enumeration proportions are relatively well estimated and the pseudo-confidence intervals are narrow. We envisage that in many sociological analyses, especially where sample and sub-sample sizes are small, using bootstrap techniques to provide pseudo-confidence intervals will be attractive.

Table 11 row 6 reports the estimated differences due to social class. We observe a negative ordinal effect consistent with the observed effects of social class. Turning our attention to young people from the other non-manual social class we note that of the original observed difference between these young people and their counterparts in the professional / managerial social class we estimate that 0.01 is due to the ‘direct’ effect of social class and 0.08 is due to their other characteristics. Sample enumeration allows a wider range of illustrations. For example if policies could be put in place that removed the direct effect of parental social class, then the observed gap of 0.20 between young people from other non-manual and skilled manual parental social class groups could be reduced to about 0.05.

In our experience we have found that by using sample enumeration it is possible to convey the effects of individual explanatory variables in an appropriate context that takes account of the other important factors identified in the modelling process. For example it is possible to convey to a policy audience, that even if we could remove the direct effects of parental social class, we couldn’t reasonably expect the same proportion of pupils in the unskilled manual social classes to attain 5+ GCSEs as those from the professional / managerial parental social class, because these pupils have other important characteristics that are different.

Estimating sample enumeration proportion provides a useful additional approach in the attempt to communicate the substantive effects of individual explanatory variables whilst recognising the effects of other variables included in the model. However, we do not claim that this approach is a panacea for communicating the results of logistic regression models. An important point to note is that sample enumeration proportions are not necessarily symmetrical. For example, if we estimate enumeration proportions for males and then for females the effect does not have to be symmetrical because individuals in these two groups will have different values for the other explanatory variables in the model. Whilst this is not a major weakness in the method it chimes with a point made by Berk (2004), that the analyst should think carefully about which aspect of the model they wish to emphasise.^{xxv} We are also aware that at the present

time there is no software that undertakes this procedure automatically. However, we hope that by making our Stata syntax files available sociologists will at least be able to follow a worked example.

Table 11 Sample Enumeration Proportions of Young People Attaining 5+ GCSEs (A*-C) Year 11 by Social Class
YCS Cohort 9 (Survey Logistic Regression Model 4)

	Professional / Managerial	Other Non-Manual	Skilled Manual	Semi-Skilled Manual	Unskilled Manual	All Classes
1 Observed proportions attaining 5+ GCSEs	0.69	0.60	0.40	0.32	0.20	0.49
<i>Sample enumeration estimates</i>						
2 Upper estimate		0.62	0.56	0.52	0.47	
3 Estimate		0.61	0.56	0.51	0.46	
4 Lower estimate		0.61	0.56	0.50	0.45	
<i>Estimates of Variable Effects</i>						
5 Observed difference with Professional / Managerial Class		0.09	0.29	0.37	0.49	
6 Estimated proportion of difference due to Social Class		0.01	0.16	0.19	0.26	
7 Estimated proportion of difference due to all other X variables		0.08	0.13	0.18	0.23	

4. Conclusion

The overall issue of interpreting and presenting the results from logistic regression models is neatly summarized by Goldstein (1993: page) who asserts that one of the useful things about statistical models is that, so long as one states the assumptions clearly and follows the rules correctly, one can obtain conclusions which are, in their own terms, beyond reproach. He continues however, by stating that the awkward thing about these models is the snares they set for the casual user, the person who needs the conclusions, but is untrained in questioning the assumptions. In Goldstein's view what makes things more difficult is that, in trying to communicate with the casual user, the data analyst is obliged to use familiar terms in an attempt to capture the essence of the model. Goldstein concludes that it is hardly surprising that such an enterprise is fraught with difficulties, even when the attempt is genuinely one of honest communication.

What follows is a summary of our recommendations, which are intended to help sociologists currently working with logistic regression models and readers of results from the models. We do not envisage that they will be the last word on the subject but hope that they will provide some useful guidelines. Our first recommendation is that sociologists should follow the advice offered in Dale (2006) and, as far as is practicable, provide as much information on the model and the model building process as possible. Access to the internet is now ubiquitous and researchers should be able to make such information available to others; if nothing else software syntax file (e.g. Stata syntax files) should routinely be made available. We believe that this will affect an important step-change and will greatly increase the potential for replication, which is integral to the incremental development of social science.

Within sociology more emphasis should be placed on data management in survey data analysis. We believe that better training in this area would remove the current obstacles and enable more substantive analyses. Texts such as those of Long (2009) are therefore highly relevant to post-graduate students and researchers. We also recommend that survey data analyst keep in contact with developments in the DAMES Node.^{xxvi}

In our experience the substantive conclusions in sociological applications from logit and probit models will generally be identical but whilst the logit model is more popular within sociology, a working knowledge of probit should also be encouraged. This is because it allows sociologists access to papers published in economics. Such knowledge is also relevant to more advanced modelling approaches for binary outcome data which tend to be developed in the probit framework, for example bivariate probit model, the maximum-likelihood probit estimator with sample selection, and the dynamic random effects model.^{xxvii} These models can all be estimated within Stata.

The development of a clear and well thought out model fitting strategy is also important in the modelling process. We recommend that sociologists should think carefully about the operationalisation and measurement of both outcome and explanatory variables. This is a general point relating to all forms of statistical model and links to our next recommendation. As we have shown, the specific

parameterisation of logistic models is consequential, especially when, as is frequently the case in sociological research, categorical explanatory variables are included in the model. The specific parameterisation can dramatically influence results in logit models and this can ensnare unwitting analysts. Sociologists should remain aware of this issue when comparing results across studies including meta analyses.

In the classical linear model, we have become familiar with the idea that a one unit change in x_1 results in a change in y equal to β_1 . The fundamental obstacle associated with logistic regression models is that there is not an obvious intuitive interpretation of the β associated with an individual explanatory variable. As we have argued this is the result of the transformation of the binary (0,1) outcome variable and the resultant estimation of β on the logit scale. Although it is not a remedy we remind readers of the useful suggestion from Gelman and Hill (2008) that as a ‘rule of convenience’ $\beta/4$ provides a means of estimating the upper bound of the predictive difference corresponding to unit change in the explanatory variable.

We strongly suggest that sociologists do not use odds and odds ratios to interpret the effects of individual explanatory variables. We have shown above that the odds scale is non-linear and this has the strong potential for providing misleading results. This problem affects substantive conclusions and is a serious problem. We would like to see the end of odds used as a mechanism to interpret logit models and we would welcome the end of these measures being reported in sociological publications.

We advocate the use of probabilities as a device to compare groups, and in our experience this has been a beneficial way of explaining the effects of explanatory variables to non-technical audiences. We have indicated that in models with a large number of explanatory variables (or categories) reporting these results can be unwieldy. Once again, depositing this information on the web is advisable. A caveat to this approach is that the researcher should be conscious of how well the model fits the data. In particular attention should be paid to the precision of the coefficients from which the probabilities are computed.

We are adamant that the computation of quasi-variances is essential to the interpretation of the effects of categorical explanatory variables in statistical models. The construction of comparison intervals from quasi-variances is critical for making appropriate comparisons that do not include the reference category. As we have demonstrated plotting coefficients and quasi-variance based comparison intervals provides a very useful graphical indication of the substantive effects of categorical explanatory variables included within logistic regression models. Therefore we believe that such graphical displays should become routine in sociological publications.

We have demonstrated that the use of sample enumeration methods provides an additional resource to help sociologists communicate the effects of individual explanatory variables. Estimating sample enumeration proportion provides the researcher with a useful means to communicate the substantive effects of individual explanatory variables whilst recognising the effects of other variables included in the model. Expressed alternatively, the method provides a means of asking ‘what if’ questions, in the multivariate context of the estimated statistical model. Because the method communicates effects in terms of proportions we have found that it is more

intuitive than expressing effects on the logit scale. This is arguably a clear benefit when communicating to non-technical audiences. We note that Payne (1999) reports the success of this method for communicating the results of voting behaviour models to Labour Party officials.

In conclusion we assert that there are clear benefits to sociologists placing more effort into interpreting the 'right hand side' of equations, by which we mean examining the specific effects of explanatory variables within models. Increased emphasis in understanding modelling results will ultimately lead to better understanding empirical regularities in the social world. We consider that logistic regression models are an essential statistical technique that is appropriate to a wide range of sociological applications which model binary outcomes. Our overall aim in this paper has been to highlight and illustrate some problems and issues associated with logit models. We believe that the recommendations and guidelines offered above are of practical value and have the potential to improve statistical modelling practices in sociological research. Whilst we do not anticipate that they will be the final word on this subject we hope that they will contribute to wider methodological discussions in the area of statistical modelling within sociology.

References

- Aldrich, J.H. and F.D. Nelson (1984) *Linear Probability, Logit, and Probit Models*. Beverley Hills, CA: Sage.
- Amemiya, T. (1981) 'Qualitative response models: A survey', *Journal of Economic Literature* 19:1483-1536.
- Berk, R.A. (2004) *Regression Analysis: A Constructive Critique*. London: Sage.
- Bulmer, M. (1996) 'The ethnic question in the 1991 census of population', in D. Coleman & J. Salt (eds) *Ethnicity in the 1991 Census, Volume 1: Demographic characteristics of the ethnic minority populations*. London: Her Majesty's Stationery Office.
- 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, J.S. (2003) *Logit models from economics and other fields*. Cambridge: Cambridge University Press.
- Cramer, J.S. (2007) 'Robustness of Logit Analysis: Unobserved Heterogeneity and Mis-Specified Disturbances', *Oxford Bulletin of Economics and Statistics* 69(4):545-555.
- 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.
- Davies, R.B. (1992) 'Sample Enumeration Methods for Model Interpretation', in P. van der Heijden, W. Jansen, B. Francis and G. Seeber (eds) *Statistical Modelling: A Selection of Papers from the Sixth International Workshop*. Amsterdam: Elsevier.
- Davies, R.B., P. Elias and R. Penn (1992) 'The Relationship between a Husband's Unemployment and His Wife's Participation in the Labour Force', *Oxford Bulletin of Economics and Statistics* 54(2):145-71.
- Department for Children, Schools and Families (2009) *GCSE and Equivalent Examinations Results in England 2007/08 (SFR 02/2009)*. London: DfCS.
- Department of Education (1985) *General Certificate of Secondary Education: A general introduction*, London, Her Majesty's Stationery Office.
- Drew, D., J. Gray, and N. Sime (1992) 'Against the Odds: The Educational and Labour Market Experiences of Black Young People', *Youth Cohort Series 19, Employment Department Training Research and Development Series*. Sheffield: Employment Department.

Elliot, B. and C. Marsh, (1988) *Exploring Data: An introduction to data analysis for social scientists (2nd Edition)*. Cambridge: Polity Press.

ESRC (2005) *Economic and Social Research Council Postgraduate Training Guidelines (4th Edition)*. Swindon: Economic and Social Research Council.

Firth, D. (2003) 'Overcoming the Reference Category Problem in the Presentation of Statistical Models', *Sociological Methodology* 33(1):1-18.

Freese, J. (2007). Replication Standards for Quantitative Social Science: Why Not Sociology? *Sociological Methods and Research* 36(2):153-171.

Gayle, V., D. Berridge and Davies, R.B. (2002) 'Young People's Entry To Higher Education: Quantifying Influential Factors', *Oxford Review of Education* 28(1):5-20.

Gayle, V., D. Berridge, and R.B. Davies, (2003) 'Econometric Analysis of the Demand for Higher Educations', *RR472 Department for Education and Skills Research Series*. Nottingham: DfES Publications.

Gayle, V. and P.S. Lambert (2007) 'Using Quasi-variance to Communicate Sociological Results from Statistical Models', *Sociology* 41(6):1191-1208.

Gayle, V., P.S. Lambert, and S. Murray, (2009) School-to-Work in the 1990s: Modelling Transitions with large-scale datasets', in Brooks, R *Transitions from Education to Work: New Perspectives from Europe and Beyond*, Basingstoke: Palgrave Macmillan.

Gelman, A. and J. Hill (2008) *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge: Cambridge University Press.

Gillborn, D. and C. Gipps, (1996) *Recent Research on the Achievements of Ethnic Minority Pupils*. London: HMSO.

Goldstein, H. (1993) 'Assessing Group Differences', *Oxford Review of Education* 19(2):131-150.

Goldthorpe, J.H. (2007) *On Sociology: Numbers, Narratives, and the Integration of Research and Theory (2nd Edition)*. Stanford: Stanford University Press.

Greene, W. H. (2003) *Econometric Analysis (5th Edition)*. Upper Saddle River, NJ: Prentice-Hall.

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

Holm, A., M.M. Jaeger, and M. Pedersen (2008) 'Unobserved Heterogeneity in the Binary Logit Model with Cross-sectional Data and Short Panels: A finite mixture approach, Research Department of Social Policy and Welfare Services, The Danish National Centre for Social Research, Working Paper 16:2008.

- Kohler, U. and F. Kreuter (2009) *Data Analysis Using Stata (2nd Edition)*. College Station, Texas: Stata Press.
- Lambert, P.S., V. Gayle, L. Tan, K. Turner, R. Sinnott, and K. Prandy, (2007) 'Data Curation Standards and Social Science Occupational Information Resources', *International Journal of Digital Curation*, 2(1)
<http://www.ijdc.net/index.php/ijdc/article/view/26/29>
- Lambert, P.S. and E. Bihagen (2007) 'Concepts and Measures: Empirical evidence on the interpretation of ESeC and other occupation-based social classifications', *International Sociological Association, Research Committee 28 Summer Meeting*, Montreal 14th-17th August.
http://www.camsis.stir.ac.uk/stratif/archive/lambert_bihagen_2007_version1.pdf
- Leckie, G. and H. Goldstein (2009) 'The limitations of Using School League Tables to Inform School Choice', *Centre for Market and Public Organisation Working Paper Series No. 09/208*.
- Leete, R. and Fox, J. (1977) 'Registrar General's Social Classes: Origins and Uses', *Population Trends*, 8:1-7.
- Levesque, R., & SPSS Inc. (2008). *Programming and Data Management for SPSS Statistics 17.0*. Chicago, IL: SPSS Inc.
- Liao, T. F. (1994) *Interpreting Probability Models – Logit, Probit and Other Generalized Linear Models*. London: Sage.
- Long, J.S. (2009) *The Workflow of Data Analysis Using Stata*. College Station, Texas: Stata Press.
- Menard, S. (1995) *Applied Logistic Regression*. London: Sage.
- Mobley, M., C. Emerson, I. Goddard, S. Goodwin, and R. Letch (1986) *All About GCSE- A clear and concise summary of all the basic information about GCSE*. London: Heinemann.
- North, J. (1987) *The GCSE: An Examination*, London: The Claridge Press.
- Payne, C. (1999) 'Helping to Win the UK Election with Statistical Models?', in H. Freidl, G. Berghold and G. Kauermann (eds) *Statistical Modelling: Proceedings of the 14th Workshop on Statistical Modelling*. Graz: University of Graz.
- SPSS (2008) *SPSS for Windows*, Release 17.0 Chicago, IL: SPSS Inc.
- Stata (2007) *Stata Release 10*: College Station, TX: StataCorp LP.
- Stewart, M. (2006) 'A Stata program for the Heckman estimator of the random effects dynamic probit model',

<http://www2.warwick.ac.uk/fac/soc/economics/staff/faculty/stewart/stata/redprobnote.pdf>

Treiman, D.J. (2009) *Quantitative Data Analysis - Doing Social Research to Test Ideas*. San Francisco. CA: Jossey-Bass.

Van de Ven, W.P. and B.M.S. Van Praag (1981) 'The demand for deductibles in private health insurance: A probit model with sample selection', *Journal of Econometrics* 17:229-252.

Endnotes

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

ⁱⁱ The UK Data Archive (www.data-archive.ac.uk) and the Economic and Social Data Service (ESDS) (www.esds.ac.uk) are emblematic. These resources have provided social scientists access to an increasing number of survey datasets. In addition these services exhibit high standard of data curation, especially in areas such as supporting documentation.

ⁱⁱⁱ Here we are referring to discrete binary outcomes, by which we mean the data can only be classified into one of two possible categories. Classically the outcome variable is either coded as 0 or 1. For example ‘not pregnant’=0 and ‘pregnant’=1.

^{iv} An argument is occasionally advanced that there is a preference for the logit model when the outcome variable is ‘truly’ discrete and the response can genuinely only be in one category, and that the probit model is appealing when the binary outcome is a ‘coarse’ grouping of an underlying, or latent, continuous variable. This view might at first be appealing, however in our experience these two models should be regarded as being two comparable alternatives which are largely mathematically equivalent.

^v One logistic regression model is fitted to the first sweep (i.e. wave) of data collected in each of the cohorts and therefore the models provide straightforward cross-sectional analyses.

^{vi} Since the early 1990s school ranking based on GCSE results have been published in league tables, initially using ‘raw’ performance measures such as the percentage of pupils gaining five or more passes at grade A*-C (Leckie and Goldstein, 2009). This measure is still published annually by The Department for Children, School and Families (see <http://www.dcsf.gov.uk/performance/tables/>).

^{vii} We envisage that with e-Social Science developments these practices will become even more widespread and will further aid replication of analyses.

^{viii} The Data Management Through e-Social Science project is a Node of the ESRC National Centre for e-Social Science (see <http://www.dames.org.uk/>).

^{ix} We are well aware that for some audiences the term ‘data management’ is associated with activities which are better thought of as being about ‘controlling’ data (e.g. tasks involved in archiving and distributing datasets, typically performed by data archivists). In this paper we are solely concerned with the issues associated with ‘preparing’ data to ‘enable’ sociological analyses.

^x SN 4009 -Youth Cohort Study of England and Wales, 1998-2000; Cohort Nine, Sweep One to Four.

^{xi} A useful short history is provided by the Qualifications and Curriculum Authority (QCA) see http://www.qca.org.uk/qca_6210.aspx (access 8.3.09).

^{xii} Survey weighted mean = 7.810; linearized standard error 0.018. This measure is computed from a derived variable deposited in SN5765 Youth Cohort Time Series for England, Wales and Scotland, 1984-2002 (Education and Youth Transitions in England, Wales and Scotland, 1984-2002).

^{xiii} Menard (1995) provides an introduction to the issue of multicollinearity in logit models. Philip B. Ender has a useful file which calculates various multicollinearity measures in Stata. This file ‘collin’ can be incorporated into the reader’s version of Stata by typing ‘findit collin’ or can be downloaded at <http://www.ats.ucla.edu/stat/stata/ado/analysis/>.

^{xiv} The standard errors reported in column 4 (and Connolly Table 5 p.20) are incorrect. More recent versions of SPSS now have a complex survey option. However, the authors have recently detected some problems with SPSS in the complex survey option when handling the complex survey design of the British Household Panel Survey. Therefore we recommend that sociologists should use Stata when analysing survey datasets.

^{xv} We are happy to provide an extended discussion of our views on sample survey weights and statistical modelling on request.

^{xvi} Two obvious examples are the Millennium Cohort Study (MCS) and the new UK Household Longitudinal Survey (Understanding Society), both of which have complex selection and sampling strategies. The MCS current has complex data analyses weights deployed with its datasets and this is the plan for the Understanding Society.

^{xvii} Treiman (2009) comments that Stata ‘has very rapidly become the statistical package of choice in leading sociology and economics departments. This is not accidental. Stata is a fast and efficient package that includes most of the statistical procedures of interest to social scientists, and new commands are added at a rapid pace’ (p.XXIV). He further asserts that ‘although it is widely used by social scientists in Europe and Asia, it has largely lost its market in leading U.S. research universities’ (p.66).

^{xviii} For a full introduction to these methods see Gayle and Lambert (2007). A set of web-based computational resources are also provided at www.longitudinal.stir.ac.uk/qv .

^{xix} See Firth (2003) for an extended statistical discussion.

^{xx} These are the three Black ethnicity categories relating to Question 58 of Sweep 1 (Cohort 9), which asks ‘Which of the following groups do you belong?’.

^{xxi} Gillborn and Gripps (1996) assert that ‘where statistics allow distinctions to be made, pupils of Black African backgrounds often achieve relatively higher results than their peers of Black Caribbean origin’ (p.34). We have re-estimated Model 2 with the black category disaggregated and find that they are not significantly different to each other. Compared with Black Caribbeans, the other two black categories are not significant, Black African (p=.968) and Other Blacks (p=.768). Black Africans are not significantly different to Other Blacks (p=.820).

^{xxii} We recommend Elliot and Marsh (2008) for introduction to the wider subject of Exploratory Data Analysis (EDA).

^{xxiii} Logit models are estimated with a fixed variance $\pi^2/3$. In the case of standard linear models we usually expect that β_1 (related to x_1) will remain consistent when another non-correlated variable x_2 is added to the model (although in practice there might be a difference this will be unimportant minor random variation). Cramer (2003) states that linear regression is unaffected by omitted variables that are not correlated with the explanatory variables included in the model, the coefficient estimates are still consistent and unbiased and the only inconvenience is a loss of precision due to increased residual variance. He continues that ‘no such comforting argument holds for logit and probit models (p.80). In a more recent paper Cramer (2007) reports more encouraging results. He asserts that in probit and logit analyses, omitting a variable will bias betas of the remaining regressors towards zero. For the probit model, Wooldridge (2002) has proved that this bias does not carry over to the partial effect of the remaining regressors or the derivatives of the outcome in their respect. Cramer asserts that for the logit model, simulations confirm that it shares this property with probit. And while omitting a variable always implies mis-specification of the disturbance, the additional effect of this on the β while significant, is generally slight, of the order of a few per cent.

We are of emerging methodological work which explores the effects of omitted explanatory variables through a latent class approach (see Holm, Jaeger and Pedersen 2008). This approach may have utility in the analysis of binary outcomes in cross-sectional data and in short panels. We suspect that such approaches may become more established in sociology in future.

^{xxiv} We use the terms log odds and logit scale interchangeably.

^{xxv} We also note that when the outcome variable is bounded, for example young people entering university, where there is a finite number of places and therefore an upper limit on the proportion an

adjustment to appropriately scale the sample enumeration proportion so that it does not exceed this upper limit is sensible.

^{xxvi} www.dames.org.uk .

^{xxvii} For discussions of the bivariate probit see Greene (2003); the maximum-likelihood probit with sample selection see Van de Ven and Van Pragg (1981); the random effects dynamic probit model see Stewart (2006).