

Statistical reanalysis of corpus data

Christer Geisler
Uppsala University

1 Introduction

A number of research problems require the analysis of a dichotomous outcome: whether a person will develop a disease, whether a child will need remedial instruction in school, or whether language users will use a particular grammatical feature. Such binary outcome variables can be analyzed using a method called logistic regression. In logistic regression, we are interested in describing the relationship between one or more so-called explanatory variables (e.g. sex, age, and occupation, as in the present study) and one dichotomous outcome variable (e.g. the choice of *that* versus *wh*-form, also in the present study).

This study reanalyzes data from a previous investigation of overt relative markers in Ulster English (Geisler 2002) with the help of logistic regression. One result of that investigation was the small proportions of *wh*-forms (such as *who*, *whom*, *whose*, and *which*), and the predominance of the relativizer *that*. Relativization in British and American English corpora has been treated in numerous studies recently. Geisler and Johansson (2001) suggest that relativization is distributed differently in British and American English: in both varieties *who* is the predominant relativizer with personal antecedents. With nonpersonal antecedents, however, the two varieties differ: in American English, speakers mainly use *that* and the relativizer *which* is more or less reserved for nonrestrictive relative clauses; in British English, both *that* and *which* are used with nonpersonal antecedents. In sum, British English is more likely to use *wh*-forms than American English. As it turns out, Ulster English is predominantly *wh*-less. *Wh*-forms are not only rare but restricted to a small group of speakers. The present report complements the previous analysis by including a number of variables in one statistical model, to provide additional insights into the variation of relativization.

2 *Corpus data*

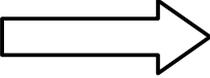
The data derive from the *Northern Ireland Transcribed Corpus of Speech* (Kirk 1990, 1997), henceforth NITC. The NITC corpus consists of interview data gathered by the *Tape-Recorded Survey of Hiberno-English Speech* and comprises over 400,000 words running text (see Adams et al. 1985). Speakers come from 38 different locations in the six counties of Northern Ireland. I will henceforth refer to the data as Ulster English, although Ulster also includes three counties in the Republic of Ireland not represented in the material.

Apart from the NITC corpus of over 400,000 words, the database compiled for this study contains sociolinguistic information about each speaker's age, age group, sex, religion, and occupation. In addition, each instance of a relative marker is coded for the form of relative marker (*that* versus *wh*-form), which serves as the outcome variable. A total of 374 relative clauses were included in the analysis. Only informant data are included in the logistic regression (excluding field-worker data from the study).

3 *Logistic regression*

The present report describes the reanalysis of the Ulster data by logistic regression. Regression is concerned with explaining the relationship between an outcome (or response) variable and one or more explanatory variables (see Table 1a). Logistic regression is used when the outcome variable is categorical. In this case, the dichotomous outcome variable has two values (*that* versus *wh*-form), which can be interpreted as a proportion between 0 and 1. A number of explanatory (or independent) variables are then submitted to the analysis. In the Ulster data, the explanatory variables include the categorical variables Sex (Male/Female), Religion (Protestant/Catholic), Occupation (where informants were categorized into Farmer, Pupil, Labourer, or White-collar worker), and the continuous variable Age (a continuous variable representing the age of the informant, ranging from 9 to 91 years of age). In short, the purpose is to analyze the relationship between this set of explanatory variables and the probability of the occurrence of *wh*-forms in the data.

Table 1a: Sets of variables in the logistic regression analysis

Explanatory variables (independent variables)		Outcome/response variable (dependent variable)
Sex, Religion, Age, Occupation		Type of relativizer (<i>wh</i> -form versus <i>that</i>)

The data was analyzed with the help of logistic regression analysis using R's *glm*-function. R is an open-source computer language for statistical computing. Good introductions to logistic regression include Demaris (1992), Hosmer and Lemeshow (2000), Pampel (2000), and Kleinbaum and Klein (2002). Apart from ample documentation at the R home page (<http://www.r-project.org>), numerous books on R exist, such as Dalgaard (2002), Venables and Ripley (2002), Verzani (2005), Sachs and Hedderich (2006), and Ligges (2007). This study will not detail the various steps used to reach the statistical results but instead focus on the output from the analyses. However, two central concepts in logistic regression are odds and logged odds. Odds express the ratio of the likelihood of an occurrence over the likelihood of a nonoccurrence (see Tables 1b and 1c). Logged odds (henceforth log. odds) are the natural logarithms of odds.¹

Table 1b: Crosstabulation of type of relative marker and sex and religion

Sex	Religion	<i>that</i>	<i>wh</i> -form	Total	Probability <i>that</i>	Probability <i>wh</i> -form	Odds of <i>wh</i> -form	Log. odds <i>wh</i> -form
Women	Catholic	40	30	70	0.57	0.43	0.75	-0.288
	Protestant	29	13	42	0.69	0.31	0.44	-0.821
Men	Catholic	155	25	180	0.87	0.13	0.15	-1.897
	Protestant	55	27	82	0.67	0.33	0.49	-0.713
Total		279	95	374	0.75	0.25	0.33	-1.099
					(1-A)	A	A/(1-A)	ln(A/(1-A))

Table 1b illustrates the relationship between raw frequencies, probabilities, odds, and log. odds. To give one example: for Catholic women, 30 instances out of a total of 70 are *wh*-forms. This corresponds to a probability of 0.43 (or 30/70). However, the odds of a *wh*-form equal 30/40 = 0.75. The natural logarithm of

these odds, or the log. odds, equals $\ln(0.75) = -0.288$. In Table 1b, frequencies of a binary variable are first transformed into a probability [A versus 1-A], then into odds $[A/(1-A)]$, and finally into log. odds $[\ln(A/(1-A))]$. As shown in Table 1c, for probabilities above 0.5, log. odds are positive, and for probabilities below 0.5 log. odds are negative (as are all the log. odds in Table 1b). Note that, when the odds equal 1 (at a probability of 0.5), the log. odds are 0.

Table 1c: The relationship between probabilities, odds, and logged odds

Probability	0.01	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.99
1-Probability	0.99	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.01
Odds	0.01	0.11	0.25	0.43	0.67	1	1.50	2.33	4.00	9.00	99.00
Log. odds	-4.60	-2.20	-1.39	-0.85	-0.41	0	0.41	0.85	1.39	2.20	4.60

In logistic regression, the effect parameters (which are labelled estimates in Tables 2 and 3) are reported in the form of log. odds, and they are generally referred to as *logits*.

The logistic regression analysis produces estimates representing log. odds for all values of the variables, together with a standardized error, a z-value (which is the estimate divided by the standardized error), and the probability associated with that z-value. In addition, Tables 2 and 3 include Wald's Chi-square statistic, which is the z-value squared, and the odds associated with each estimate (following the guidelines in Peng *et al.* 2002). Since the estimates in Tables 2 and 3 represent log. odds, the odds for a variable are simply the estimate x (representing the natural logarithm of the odds) raised to the power of e (e^x). For example, in Table 2, the odds of a *wh*-form among women informants are $e^{0.2198} = 1.25$.

The logistic regression analysis was carried out in two steps. First, the whole set of explanatory independent variables was included (Sex, Religion, Age, and Occupation): this is referred to as model 1 (Table 2). In a second step, only those variables that were considered to be statistically significant were kept in the second analysis; this second analysis is referred to as model 2 (Table 3). Finally, predicted odds and probabilities for all combinations of the categorical variables were calculated to show the practical applicability of logistic regression (Table 4). The analyses in Tables 2 and 3 only include so-called main effects; that is, possible interaction effects between the explanatory variables have not been included in this presentation (cf. Hosmer and Lemeshow 2000: 31–44).

Table 2: List of estimates from the first logistic regression (model 1)

	Estimate	Standard error	z value	Wald's χ^2	Probability (> z)	Odds
(Intercept)	1.4750	0.7166	2.06	4.24	0.04	
(Sex)=Female	0.2198	0.3137	0.70	0.49	0.48	1.25
(Religion)=Catholic	-0.3455	0.2939	-1.18	1.39	0.24	0.71
Age	-0.0639	0.0134	-4.77	22.75	< 0.01	0.94
(Occupation)=Labourer	-0.1281	0.4682	-0.27	0.07	0.78	0.88
(Occupation)=Pupil	-2.4701	0.6727	-3.67	13.47	< 0.01	0.08
(Occupation)=White-collar	1.5191	0.3849	3.95	15.60	< 0.01	4.57

Null deviance: 423.89 on 373 degrees of freedom
 Residual deviance: 313.50 on 367 degrees of freedom
 AIC: 327.50 Log. likelihood: -156.75 (df=7)

In Table 2, the column labelled Estimate provides the predicted estimates which reflect the effects of the variables on the log. odds of *wh*-forms in the data. The values of the estimates are interpreted as decrements or increments to the log. odds on that variable. The intercept (marked as Intercept in Tables 2 and 3) represents a baseline log. odds for all variables equal to 0, that is, a male, Protestant farmer at (the meaningless) age of 0. Of course, no such individual exists, but see the discussion of Age in section 4 below. The logistic regression model shows the changes in log. odds in a one-unit change in the independent variables: Sex, Religion, and the various levels of Occupation all have a one-unit change, namely from 0 to 1. For the variable Sex, Male is coded 0, and Female is coded 1. For the variable Religion, Protestant is coded 0, while Catholic is coded 1. For the variable Occupation, three so-called design variables (or dummy variables) are set up. It is important to know what values of the explanatory variables are marked as 0, since these form the baseline categories against which the logistic regression estimates are compared.

To interpret the actual estimates in Table 2, we find that being female is estimated to raise the log. odds by 0.2198, compared with what would be expected if there were no statistical association between sex and choice of relative marker. In the same way, being Catholic lowers the log. odds by -0.3455. The estimate for the continuous variable Age shows that a one-year increase in age decreases the log. odds of a *wh*-form by -0.0639. The last variable, Occupation,

has four levels, and the three levels in Table 2 show that the baseline category is Farmer, against which all other levels of Occupation are compared. Hence, being a pupil at school decreases the log. odds by -2.47, while being a white-collar worker increases the log. odds by 1.5191. One problem with log. odds is that they lack a meaningful interpretation. Instead, log. odds can be transformed into odds. An odds is defined here as the ratio of the probability of using a *wh*-form over the probability of using the relativizer *that*.

An alternative way of interpreting the estimates in Tables 2 and 3 is to transform each estimate into odds. This is done by exponentiating each estimate. For women informants the log. odds estimate equals 0.2198 and the odds of a *wh*-form among women equal $e^{0.2198} = 1.25$. That is, the odds of a *wh*-form are 1.25 times higher for women compared with men. In other words, the odds are 1.25 to 1 for women (odds are multiplicative). The highest odds are among white-collar workers with $e^{1.5191} = 4.57$, and the lowest odds are found among pupils with $e^{-2.4701} = 0.08$. The size of the odds can also be expressed as a percentage change in the odds (by subtracting 1 and multiplying by 100) (see Pampel 2000: 22–23). The odds increase by 25 percent among women $[(1.25 - 1) * 100 = 25]$. Being Catholic lowers the percentage change in the odds by 29 percent $[(0.71 - 1) * 100 = -29]$. Similarly, for the continuous variable Age, the estimated log. odds equal -0.0639 and the odds equal $e^{-0.0639} = 0.94$. As a percentage change in the odds, there is a 6 percent decrease in the odds of a *wh*-form for a one-year increase in age $[(0.94 - 1) * 100 = -6]$.

In the first analysis (model 1), only two variables are flagged as statistically significant: Age and Occupation. The z-values of Age and Occupation have probabilities less than 0.05 (this is also shown by Wald's Chi-square statistic in Table 2). Two variables, Sex and Religion, have z-values and Wald's Chi-square values with probabilities above 0.05. In other words, only two out of the original four variables in model 1 are statistically different from zero. A second analysis is then carried out with only the two remaining variables, Age and Occupation (Table 3).

Table 3: List of estimates from the second logistic regression (model 2)

	Estimate	Standard error	z value	Wald's χ^2	Probability (> z)	Odds
(Intercept)	1.4375	0.6716	2.14	4.58	0.03	
Age	-0.0665	0.0129	-5.15	26.52	< 0.001	0.94
(Occupation)=Labourer	-0.1584	0.4634	-0.34	0.12	0.73	0.85
(Occupation)=Pupil	-2.5226	0.6694	-3.77	14.21	< 0.001	0.08
(Occupation)=White-collar	1.5483	0.3753	4.13	17.05	< 0.001	4.70

Null deviance: 423.89 on 373 degrees of freedom

Residual deviance: 315.28 on 369 degrees of freedom

AIC: 325.28 Log. likelihood: -157.64 (df=5)

Table 3 shows the results of the second reduced model 2, where the two statistically non-significant variables Sex and Religion have been removed. A likelihood-ratio test between model 1 and 2 indicates that model 2 is preferred, since the elimination of the two variables Sex and Religion is not statistically significant: the likelihood-ratio test between model 1 and 2 = $-2 * [(-156.75) - (-157.64)] = 1.79$ (df = 2), $P[\chi^2(2) > 1.79] = 0.59$. Hence, the reduced model with only Age and Occupation is preferred.

Figure 1 shows the conditional probabilities of *wh*-forms (dashed line) across Age and the predicted probabilities of *wh*-forms across the four levels of Occupation. Informants in the lower middle age band have higher probabilities of *wh*-forms than older informants. In addition, white-collar workers have considerably higher predicted probabilities of *wh*-forms than the other occupational groups.

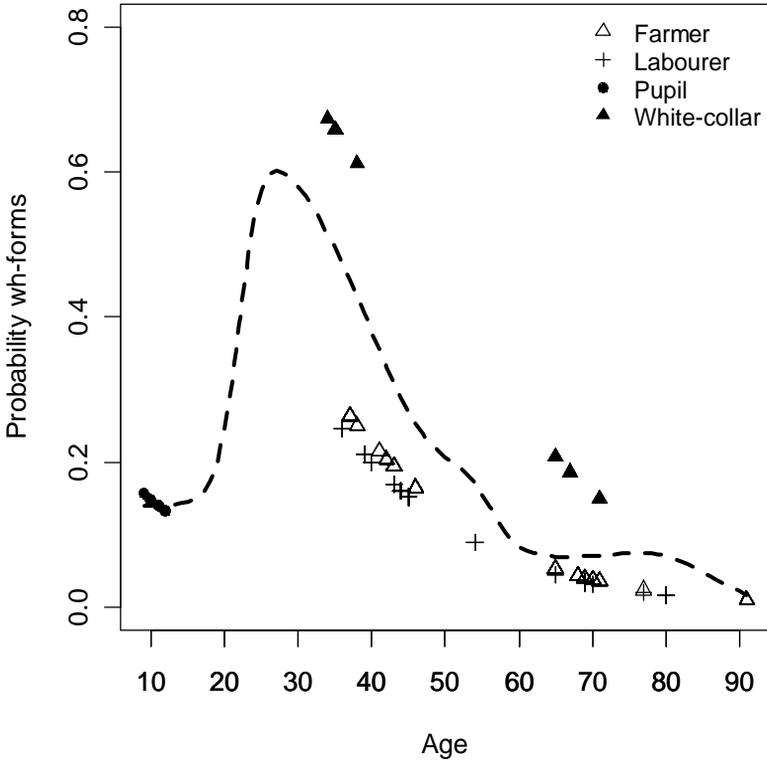


Figure 1: Predicted probabilities of wh-forms across age and occupation

4 Predicting odds and probabilities

One way of understanding the output from a logistic regression analysis is to calculate odds and probabilities for each combination of the categorical variables in the data set. This is shown in Table 4: the odds and probabilities are calculated based on the estimates in model 1 (see Table 2).

Table 4 gives the odds and probabilities for all combinations of the categorical variables at two age levels in the data set. In order to make sense of the Age variable, it is set to 30 and 60 years of age in Table 4. I am aware that two cells include 30- and 60-year-old pupils: however, we are dealing here with a statistical model.

Table 4: Predicted odds and probabilities for informants aged 30 and 60

Sex	Religion	Occupation	Odds at Age=30	Probability at Age=30	Odds at Age=60	Probability at Age=60
Male	Protestant	Farmer	0.643	0.391	0.095	0.086
Male	Protestant	Labourer	0.565	0.361	0.083	0.077
Male	Protestant	Pupil	0.054	0.052	0.008	0.008
Male	Protestant	White-collar	2.936	0.746	0.432	0.302
Male	Catholic	Farmer	0.455	0.313	0.067	0.063
Male	Catholic	Labourer	0.400	0.286	0.059	0.056
Male	Catholic	Pupil	0.038	0.037	0.038	0.037
Male	Catholic	White-collar	2.078	0.675	0.306	0.234
Female	Protestant	Farmer	0.801	0.445	0.118	0.105
Female	Protestant	Labourer	0.704	0.413	0.104	0.094
Female	Protestant	Pupil	0.068	0.063	0.010	0.010
Female	Protestant	White-collar	3.658	0.785	0.538	0.350
Female	Catholic	Farmer	0.567	0.362	0.083	0.077
Female	Catholic	Labourer	0.499	0.333	0.073	0.068
Female	Catholic	Pupil	0.048	0.046	0.007	0.007
Female	Catholic	White-collar	2.589	0.721	0.381	0.276

Two examples are explained below. First, the odds of a *wh*-form in the group Male, Protestant, Farmer, aged 30 are calculated as follows: $e^{1.47+0+0+(-0.0639*30)+0} = e^{-0.442} = 0.643$. In other words, for a particular combination of variables, the various estimates in the model are simply added together and then exponentiated (the logistic regression estimates are additive). Note the value 0 for three of the variables, namely Sex, Religion, and Occupation: $e^{1.47+0(\text{Sex}=0)+0(\text{Religion}=0)+(-0.0639*30)+0(\text{Occupation}=0)}$, since these were all coded as 0 in the data set. The estimates used for the calculations can be found in Table 2. Next, using the odds, we can also calculate a predicted probability for a particular group. The probability of a *wh*-form in this group (male, Protestant, Farmer, aged 30) equals $[0.643/(1 + 0.643)] = 0.391$. As one additional example, the odds of a *wh*-form in the group Female, Catholic, White-collar worker, aged 30 are calculated in the same way as above: $e^{1.47+0.2198+(-0.3455)+(-0.0639*30)+1.5191} = e^{0.9514} = 2.589$. The predicted probability for this group is calculated as follows: $[2.589 / (1 + 2.589)]$

= 0.721. Hence, among informants aged 30, male Protestant farmers have approximately 39 percent *wh*-forms, while female Catholic white-collar workers have about 72 percent *wh*-forms. Figure 1 shows that this approximation makes sense: white-collar workers around 30 years of age have high probabilities of using *wh*-forms.

Table 4 shows that the probabilities of *wh*-forms decrease with age: the probabilities for 60-year-olds are consistently lower than for 30-year-olds. Moreover, comparisons between the probabilities show that women have higher probabilities than men, and Protestants have higher probabilities than Catholics, while white-collar workers have higher probabilities than any other occupational group.

5 Conclusion

Logistic regression is used extensively in various disciplines, such as epidemiology and biomedical research, because of the possibility of interpreting the effects of explanatory variables on the relative risk of outcomes such as presence of a disease. Logistic regression estimates provide a simple summary of the influence of a variable on the log. odds of having a certain characteristic. Log. odds can easily be transformed into odds, which in turn can be transformed into probabilities. This study shows that logistic regression can also be used in the analysis of corpus data. The purpose of the analyses was to uncover the relationships between four explanatory variables (Sex, Religion, Age, and Occupation) and a binary outcome variable, namely type of relativizer. Out of the four explanatory variables, only two were found to be statistically significant: Age and Occupation. It was shown that the elimination of two of the variables, Sex and Religion, had no statistically significant contribution to a second reduced logistic model. In a final step in the interpretation of the data, predicted odds and probabilities of *wh*-forms were calculated for two different age groups.

Note

1. The natural logarithm is the logarithm to the base e , where e is approximately equal to 2.7182818. The natural logarithm of a number x is the power to which e would have to be raised to equal x . In section 3, when we calculate the odds of an estimate, the base e is raised to the power of the estimate. For the log. odds of -4.60 in the first column of Table 1c, the odds of 0.01 can be obtained by raising the base e to the log. odds of -4.60: $e^{-4.60} = 0.01$, which is equal to the ratio of the two proportions 0.01/0.99. Hence $\ln(0.01) = -4.60$, and $e^{-4.60} = 0.01$.

References

- Adams, George Brendan, Michael V. Barry and Philip M. Tilling. 1985. The tape-recorded survey of Hiberno-English speech. In J. M. Kirk, S. Sanderson and J. D. A. Widdowson (eds.). *Studies in linguistic geography – the dialects of English in Britain and Ireland*, 67–80. London: Croom Helm.
- Dalgaard, Peter. 2002. *Introductory statistics with R*. New York: Springer.
- Demaris, Alfred. 1992. *Logit modeling: Practical applications*. Newbury Park, CA: Sage.
- Geisler, Christer. 2002. Relativization in Ulster English. In P. Poussa (ed.). *Relativisation on the North Sea Littoral* (LINCOM Studies in Language Typology 07), 135–146. München: Lincom Europa.
- Geisler, Christer and Christine Johansson. 2001. Relativization in formal spoken American English. In M. Modiano (ed.). *Studies in Mid-Atlantic English* (HS-institutionens skriftserie No. 7), 87–109. Gävle: Gävle University.
- Hosmer, David W. and Stanley Lemeshow. 2000. *Applied logistic regression*. Second edition. New York: Wiley.
- Kirk, John M. 1990. *Northern Ireland Transcribed Corpus of Speech*. Colchester: University of Essex.
- Kirk, John M. 1997. Ulster English: The state of the art. In H. Tristram (ed.). *The Celtic Englishes*, 135–176. Heidelberg: C. Winter.
- Kleinbaum, David G. and Mitchel Klein. 2002. *Logistic regression. A self-learning text*. Second edition. New York: Springer.
- Ligges, Uwe. 2007. *Programmieren mit R*. Second edition. Berlin: Springer.
- Pampel, Fred C. 2000. *Logistic regression. A primer*. Thousand Oaks, CA: Sage.
- Peng, Chao-Ying Joanne, Kuk Lida Lee and Gary M. Ingersoll. 2002. An introduction to logistic regression analysis and reporting. *Journal of Educational Research* 96: 3–14.
- Sachs, Lothar and Jürgen Hedderich. 2006. *Angewandte Statistik. Methodensammlung mit R*. 12th edition. Berlin: Springer.
- Venables, William N. and Brian D. Ripley. 2002. *Modern applied statistics with S*. Fourth edition. New York: Springer.
- Verzani, John. 2005. *Using R for introductory statistics*. Boca Raton, FL: Chapman and Hall.

