

# **Binary Logistic Regression and its application to data from a study of children's recognition of their own recorded voices**

(Term paper in Statistical Methods, Spring 2009)

Sofia Strömbergsson

## **Introduction**

This paper describes a statistical method that I've used when analyzing the data in a study of children's ability to recognize their recorded voice as their own. To begin with, I had no real understanding of the method I was applying, but simply inserted my data into the formula and computed the results. As I was not completely comfortable that my selection of method was appropriate, I felt a need for a deeper understanding of what I was doing. This paper is the result of my efforts to gain this deeper understanding.

## **Data description**

The data is from a study of children's ability to identify which of 4 recordings is their own voice. 45 children participated as subjects. 3 children were recorded as references. For each word in the recording script (23 in total), the task for the subjects was to

- 1) Record their own production of the word
- 2) Decide which one of four recordings of this word (their own recording + 3 reference recordings, presented in random order) is their own voice.

There were two test occasions; on the first, the children performed both the recording and the identification task, on the second, they only performed the identification task.

The recordings (23 words \* 45 subjects = 1035 recordings) represent observations, and for each recording the following information is specified:

- The number of phonemes in the word (ranging from 2 to 5)
- The absolute difference (in Hertz) between the subject's average F0 and the most similar reference speaker's average F0
- The Euclidean difference between the subject's average formant frequencies and the most similar reference speaker's average formant frequencies
- The absolute difference between the subject's speaking rate (in phonemes/sec) and the most similar reference speaker's speaking rate
- One test result (correct/incorrect identification) for each subject on the first test occasion
- One test result (correct/incorrect identification) for each subject on the second test occasion

## Statistical analysis

The following paragraphs present a general introduction the binary logistic regression. The last paragraph in this section will describe how the logistic regression was applied to the data described above.

### *Multiple regression and ordinary least squares (OLS) estimation*

Before exploring the logics behind the logistic regression, it can be helpful to look at linear regression models. Regression models are used to explore the relationship between a dependent variable  $Y$  and one or multiple independent variables  $X_1$  to  $X_n$ ; if  $Y$  varies systematically with different values for  $X_1$  to  $X_n$ . The model is usually described by the following formula (Spicer, 2004):

$$Y = [\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon] + r$$

where  $\varepsilon$  is a constant representing the random error (any other effects than  $X_1$  to  $X_n$ ),  $\beta_0$ ,  $\beta_1$  and  $\beta_n$  are regression coefficient parameters (the amount  $Y$  changes by each one-unit change in  $X_x$ , while other independents are held constant). Each regression coefficient  $\beta$  represents the slope of the regression line; the larger the  $\beta$ , the more influence the independent variable  $X$  has on the dependent variable  $Y$ .  $r$  is the residual, i.e. the discrepancies between the predicted  $Y$  values and the actual  $Y$  values. A common way to calculate the regression coefficients is to use ordinary least squares (OLS) estimation. Here, values for the regression coefficients  $\beta_x$  and  $\varepsilon$  are chosen that minimize  $r$ , i.e. minimizing the summed squared differences between all predicted  $Y$  values and the corresponding actual  $Y$  values.

Hypothetically, if this model was applied to data where the dependent variable  $Y$  were categorical (as is the case for my data), the composite variable in the square brackets above would no longer predict scores, but *probabilities* of a case being in the category labeled 1. So, each regression coefficient  $\beta_x$  would then be the change in probability of a case belonging to category 1, given a 1-unit increase in the independent variable  $X_x$ . Although this might well be interpretable and might seem well-suited, there are several reasons why OLS regression should not be used when dealing with dependent variables that are categorical. One important reason is that the generated probabilities may fall outside the 0 to 1 range (Spicer, 2004). Instead, binary logistic regression is the recommended method.

### *Binary Logistic Regression*

Binary logistic regression is the method of choice if your dependent variable is binary (dichotomous) and you wish to explore the relative influence of continuous and/or categorical independent variables on your dependent variable, and to assess interaction effects between the independent variables. The following paragraphs are based on Spicer's (2004) nice and comprehensible introduction to binary logistic regression.

As in the hypothetical case described above, where OLS regression was applied to data with a binary dependent variable, the binary logistic regression also generates predicted probabilities of a case being in the category labeled 1. However, in the binary logistic

regression, probability is expressed as *odds*. If the probability of belonging to category 1 is 80/100 = 80%, the odds of belonging to category 1 is 4. In other words, a case is 4 times more likely to be in category 1 than in category 0. (The probability is the odds/(odds + 1).) Moreover, the odds is transformed into log odds, *logits*, which is the natural log of the odds, i.e. the power to which 2.718 has to be raised to produce the odds. These transformations solve the problems that OLS regression faces when applied to data where the dependent variable is binary. Described quasi-formally, what we have now is:

$$\text{Predicted logit of } Y = [\beta_1 X_1 + \beta_2 X_2 + \dots + \varepsilon]$$

(Note that  $\beta$  are now *logistic* coefficients.) In order to find the optimal coefficient values, i.e. the values that maximize the predictive power of the coefficients, a *maximum likelihood* criterion is applied. The goal is still to find the coefficients that will produce the logits (and thereby the predicted probabilities) that will most accurately place cases in their actual category, i.e. to minimize the discrepancy between a case's predicted category and its actual category. A *log likelihood function* is used to collect the predicted probability and the actual category for all cases, given different coefficient values. The coefficients that maximize the value of this function are the ones that are finally selected as the logistic coefficients.

#### *Data requirements and related methods*

For data where the dependent variable is categorical (usually dichotomous), and all independent variables are categorical, or if they are a mix of continuous and categorical variables, or where the independent variables are all continuous but not normally distributed, logistic regression is recommended (Wuensch, 2004). In cases where the independent variables are all continuous and nicely distributed, *discriminant analysis* is often employed (Wuensch, 2004). If the dependent variable had been quantitative continuous (instead of binary as in my data), a One-Way Analysis of Variance (*One-Way ANOVA*) would have been appropriate to use (Page et al, 2003).

The sample size required to perform reliable estimations with the binary logistic regression method is greater than that needed for OLS regression, and varies with the number of independent variables. The more independent variables you have, the more cases are needed. However, the recommended minimum of cases per independent variable varies; Spicer (2004) refers to a recommendation of 50 cases per independent variable, while Garson (2009) refers to recommendations of 10 cases per independent variable. (Obviously, these recommendations are not applicable to continuous variables.)

#### *Reporting results*

Peng et al (2002) recommend that the data reported on the results of the logistic regression analysis should include

- 1) An overall evaluation of the model

The overall evaluation demonstrates if the logistic model fits the data closer than the intercept-only model (i.e. the model where the independent variables don't influence the dependent variable). In other words, does knowledge of the independent variables improve our ability to predict outcome (the value of the

dependent variable), compared to merely guessing that all cases fall into the most common outcome category. This can be examined by e.g. likelihood ratio tests and score tests (Peng et al, 2002). p-values smaller than 0.05 indicate that the independent variables most likely influence the dependent variables.

- 2) Statistical tests of individual predictors (independent variables)  
Statistical significance of individual predictors is tested using the Wald chi-square statistic. Those predictors whose p-values are smaller than 0.05 are significant.
- 3) Goodness-of-fit test statistics  
The goodness-of-fit tests indicate the appropriateness of the model, how well it fits with the actual outcomes. This can be estimated with the Hosmer-Lemeshow test, where the insignificance of the  $\chi^2$ -value is an indicator of goodness-of-fit. ( $p > 0.05$  indicates that the model fits the data well.)
- 4) An assessment of the predicted probabilities  
The predictive accuracy of the model can be presented in a classification table, where the predicted outcome (1/0) is compared to the actual outcome (1/0). The classification table is especially recommended in a report if classification is a stated goal of the analysis (Peng et al, 2002: 8). According to Garson (2009), however, the Hosmer-Lemeshow  $\chi^2$ -test of goodness-of-fit is often preferred over classification tables.

According to Spicer (2004), many reports of logistic regression analysis don't include information about the  $\chi^2$  for the model (3), classifications results (4), and the logistic coefficients, but instead focus on the odds for each independent variable and their significance.

#### *Applying the Binary Logistic Regression to the data*

In order to explore if the result (whether or not the subjects are able to identify their own recorded voice correctly) is influenced by the number of phonemes in an item (a word) and/or the acoustic similarity (in speaking rate, F0 and formant frequencies) between the subject's production and the references' production of the word, the SPSS logistic regression tool was used to compute a binary logistic regression.

Two separate logistic regressions were performed; one with the result from the first test occasion as the dependent variable, and the other with the result from the second test occasion as the dependent variable. In both models, the independent variables were

- number of phonemes in the item (NUM\_PHONEMES)
- the distance between a subject's speaking rate in the production of the item and the most similar reference's speaking rate (RATE\_DIST)
- the difference between a subject's average F0 in the production of the item and the most similar reference's F0 (F0\_DIST)
- the between a subject's average formant frequencies in the production of the item and the most similar reference's formant frequencies (FORMANT\_DIST)

## Results

The results from the logistic regression with the first test result as the dependent variable are presented in Table 1-4. As shown in Table 1, the only significant (but weak) predictor of the result on the first test occasion is FORMANT\_DIST, i.e. spectral similarity ( $p < 0.05$ ). Thus, with every unit increase of the spectral difference between a subject's recording and the reference recordings, the odds that the subject's answer will be correct increase by a factor of 1.001.

**Table 1: Logistic Regression Analysis of children's identification of their recorded voices: immediately after recording**

	B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)						
RATE_DIST	-,015	,083	,033	1	,855	,985
F0_DIST	,003	,003	1,128	1	,288	1,003
FORMANT_DIST	,001	,000	6,302	1	,012	1,001
NUM_PHONEMES	,015	,091	,026	1	,872	1,015
Constant	1,046	,384	7,429	1	,006	2,845

a Variable(s) entered on step 1: RATE\_DIST, F0\_DIST, FORMANT\_DIST, NUM\_PHONEMES.

The only model evaluation that SPSS generates is performed by measures of log likelihood (presented in Table 2). Both Peng et al (2002) and Spicer (2004) warn against relying on log likelihood measures in logistic regressions, but Spicer (2004) suggests that the two different  $R^2$  measures might be viewed as tentative indicators of the range within which the actual influence of the independent variables on the dependent variable lies. For this data, the independent variables would thus explain somewhere between 0.9% and 1.4% of the variation in results. But, Spicer still advises cautious use of these "pseudo statistics"; "they are best treated with caution if not actually avoided" (Spicer, 2004: 129).

**Table 2: Model Summary of the logistic regression analysis of children's identification of their recorded voices: immediately after recording**

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	964,528(a)	,009	,014

a Estimation terminated at iteration number 4 because parameter estimates changed by less than ,001.

The goodness-of-fit of the model is displayed in Table 3. The non-significant  $\chi^2$  ( $p > 0.05$ ) indicates that the model fits the data well.

**Table 3: Hosmer and Lemeshow Test of goodness-of-fit of the logistic regression analysis of children's identification of their recorded voices: immediately after recording**

Step	Chi-square	df	Sig.
1	8,531	8	,383

The classification table (Table 4) displays the agreement between predicted and actual results. This table reveals that the incorrect answer is never predicted. This would be the same as the intercept-only model, without independent variables, where the probability of

a correct answer is equal to <the number of correct answers>/<the total number of answers>, which in this case is 0.82.<sup>1</sup>

**Table 4: Classification Table(a) of the logistic regression analysis of children’s identification of their recorded voices: immediately after recording**

Observed			Predicted		Percentage Correct
			RL_CORRECT10		
			0	1	0
Step 1	RL_CORRECT10	0	0	186	,0
		1	0	846	100,0
	Overall Percentage				82,0

a The cut value is ,500

The results from the logistic regression with the second test result as the dependent variable are presented in Table 5-8. Here, both F0\_DIST and FORMANT\_DIST are significant (but weak) predictors of the result (p < 0.05).

**Table 5: Logistic Regression Analysis of children’s identification of their recorded voices: 1-2 weeks after recording**

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	RATE_DIST	,084	,076	1,229	1	,268	1,088
	F0_DIST	,013	,003	24,365	1	,000	1,013
	FORMANT_DIST	,001	,000	7,281	1	,007	1,001
	NUM_PHONEMES	,054	,079	,461	1	,497	1,055
	Constant	-,005	,334	,000	1	,988	,995

a Variable(s) entered on step 1: RATE\_DIST, F0\_DIST, FORMANT\_DIST, NUM\_PHONEMES.

The “pseudo statistics” in the evaluation of the model as presented in Table 6, indicates that the independent variables explain somewhere between 3.9% and 5.6% of the variation in the dependent variable. Again, caution is advised when dealing with log likelihood statistics (Spencer, 2004; Peng et al, 2002).

**Table 6: Model Summary of the logistic regression analysis of children’s identification of their recorded voices: 1-2 weeks after recording**

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1181,109(a)	,041	,059

a Estimation terminated at iteration number 4 because parameter estimates changed by less than ,001.

<sup>1</sup> This might seem surprising, as the Omnibus tests of Model Coefficients, also generated by SPSS, present significant p-values, indicating that the model containing the independent variables is significantly different from the intercept-only model (Garson, 2009). However, as the coefficients are generally very small (see Table 1), their influence on the dependent variable is weak – so weak that the changes in the predicted p-values are not reflected in the Classification table. Predictions of events with the actual outcome 1 might have come closer to 1, and predictions of events with the actual outcome 0 might have come closer to 0 – but still without falling below the cutoff value of 0.5.

The goodness-of-fit of the model is displayed in Table 7. The non-significant  $\chi^2$  ( $p > 0.05$ ) indicates that the model fits the data well.

**Table 7: Hosmer and Lemeshow Test of goodness-of-fit of the logistic regression analysis of children's identification of their recorded voices: 1-2 weeks after recording**

Step	Chi-square	df	Sig.
1	7,129	8	,523

The classification table (Table 8) displays the agreement between predicted and actual results. As was the case with the results from the first test occasion, Table 8 also reveals that the incorrect answer is actually never predicted.

**Table 8: Classification Table(a) of the logistic regression analysis of children's identification of their recorded voices: immediately after recording**

Observed			Predicted		Percentage Correct
			L_CORRECT01		
			0	1	0
Step 1	L_CORRECT01	0	0	289	,0
		1	0	743	100,0
	Overall Percentage				72,0

a The cut value is ,500

## Discussion

### *The choice of statistic analysis*

Considering that the dependent variables are binary in the data I'm analyzing, and that the independent variables are continuous but not normally distributed, I'm now comfortable that the binary logistic regression is an appropriate choice of statistical analysis. Although there are mathematical details in the method that I don't fully grasp (e.g. why is the natural log used, how does the maximum likelihood estimation operate etc.), this level of understanding is what I aimed for within the framework of this course.

The classification tables (Table 4 and 8) don't add very much to the reporting of results, part from potential confusion. As classification is not a stated goal in my study (as it is in e.g. studies of medical diagnoses or admittance to schools or universities, and how these can be predicted from different factors), the classification tables are probably best left outside of reporting the results (Peng et al, 2002: 8). For my data, the results from the Hosmer-Lemeshow  $\chi^2$ -tests are probably better illustrations of goodness-of-fit (Garson, 2009).

Although I presented log likelihood statistics of the predictive influence of the independent variables on the outcome in this report (Tables 2 and 6), I lean towards not including them in the scientific report of the results, as they are only "pseudo statistics"

(Peng et al, 2002; Spicer, 2004). What is left to present is the figures in Tables 1 and 5, together with the results of the Hosmer-Lemeshow  $\chi^2$ -tests (Tables 3 and 7).

### *Interpreting the results*

The results generated by means of the binary logistic regression indicate that similarity in F0 and formant frequencies between one's own voice and three reference voices seem to make the task of identifying which one of these four recordings represents one's own voice more difficult, at least when the identification task is not performed directly after the recording. As F0 and formant frequencies are interrelated, it should not be surprising that they both influence the children's performance. F0 similarity does not influence the children's performance when the identification task is performed immediately after recording, while spectral similarity seem to make identification slightly more difficult. Children's performance is not influenced by similarity in speaking rate between the own recorded voice and the reference voices, or by the length of the stimuli (measured in phonemes). A t-test (not reported here) revealed that the children performed significantly worse on the second test occasion than on the first. These findings indicate that children use different acoustic cues when discriminating their own voice from those of others when they perform the identification task immediately after recording, compared to when the identification task is performed after a time span of 1-2 weeks.

### **References**

- Garson, G. D. (2009). "Logistic Regression" from *Statnotes: Topics in Multivariate Analysis*. Retrieved 6/5/2009 from <http://faculty.chass.ncsu.edu/garson/pa765/statnote.htm>.
- Page, M. C., Braver, S. L. & MacKinnon, D. P. (2003). *Levine's guide to SPSS for analysis of variance*. Lawrence Erlbaum Associates, New Jersey.
- Peng, Chao-Ying Joann; Lee, Kuk Lida; & Ingersoll, Gary M. (2002). An introduction to logistic regression analysis and reporting. *Journal of Educational Research* 96(1): 3-13.
- Spicer, J. (2004) *Making sense of multivariate data analysis*. Sage Publications, California, 123-151.