REQUEST FOR A COPY

Request ID:

Request ID: 19358590 (OCLC:ORE)
Request date: 4/17/2006 **Need before:** 5/17/2006 **Max cost:** 21.25(USD)
Will pay fee: Yes **Reciprocal agreement:** **Payment provided:**
Copyright Compliance: US:CCG **Required format type:** **Payment type:** ~~IFM~~ IFM
Deliver via: ARIEL: **Level of service:** **Expiry date:** 4/20/2006

Call number:
Author:
Title: Understanding statistics.
Edition:
Series and number:
Published: Mahwah, N.J. : Lawrence Erlbaum Associates, c2002-,

Author of excerpt: Nick Sofronicu‌: Confidence Intervals for the Predictions of Logistic Regression in the Presence and Absence of a Variance- Covariance Matrix
Volume/issue: 1 1 (2002)
Pagination: 3-18
ISBN: **ISSN:**
Other ID numbers: System #:CAN:46804256
Verification source: <TN:189999> OCLC

Requester notes: Requester Number--(OCLC) ILLNUM:19358590; FAX/ARIEL:541-737-1328 EMAIL:valley.ill@oregonstate.edu

Responder name:
Responder ID: OCLC:PAU

University of Pennsylvania Libraries
Interlibrary Loan
3420 Walnut St.
Philadelphia, PA 19104-6277
USA

Date sent: **Sent via:** **Charges:**
# of pages:

Responder notes:

Requester name:
Requester ID: OCLC:ORE **Account number:**

Ship to:
Library-ILL/Oregon State University/121 The Valley Library/Cross Streets Jefferson Way & Waldo Pl/Corvallis, OR 97331-4501

Bill to:
same.O005911 ** GWLA MEMBER **

Deliver via: ARIEL: OSU-ILL.library.orst.edu when possible

Client name: Bulatov, Yaroslav
Status: **ID:**

ariel

Transaction ID: OCLC:ORE-19358590:PAU

UNDERSTANDING STATISTICS, 1(1), 3–18

RESEARCH ARTICLES

Confidence Intervals for the Predictions of Logistic Regression in the Presence and Absence of a Variance–Covariance Matrix

Nick Sofroniou
Educational Research Centre
Saint Patrick's College, Dublin

Graeme D. Hutcheson
Faculty of Education, Research, and Graduate School
University of Manchester

Confidence intervals for fitted values provide valuable information about the usefulness of regression models. Although such intervals can be easily calculated using standard statistical software for response variables that have normally distributed errors (e.g., in ordinary least-squares regression), it is more difficult to calculate them for response variables that have binomially distributed errors (e.g., logistic regression). Although a number of statistical packages provide confidence intervals for fitted values directly for logistic regression models, some commonly used packages do not (e.g., SPSS). In this article we outline a method of calculating these intervals simply by fitting a model after transforming variables. This technique is evaluated by comparing results with those obtained using a method that utilizes the variance–covariance matrix. Both techniques are described in detail and applied to simple and multiple logistic regression along with step by step instructions and software commands for SPSS Version 10.1.

Requests for reprints should be sent to Graeme D. Hutcheson, Faculty of Education, Research and Graduate School, University of Manchester, Humanities Building, Oxford Road, Manchester, M13 9PL England. E-mail: mewssgdh@man.ac.uk

In logistic regression, a form of maximum likelihood estimation is used to obtain parameters that make the observed results most likely for a response variable with binomial errors (Agresti, 1990; Collett, 1991). These parameters are calculated for changes on the logit(p) scale, the log odds of the probability p of the chosen response alternative, that is, log $(p/(1-p))$.

The relation between an explanatory variable x fitted to the model and logit(p) is a straight line, as in Equation 1.

$$\text{logit}(p) = \alpha + \beta x. \tag{1}$$

The parameters of the model are therefore interpreted in much the same way as they are in ordinary least-squares regression with the gradient, or slope of the line, indicated by the parameter β that is interpreted as the change in logit(p) resulting from a unit change in x. Put another way, β represents the change in the log odds of an event happening for a unit change in x. The probability that a particular event will occur may then be calculated using Equation 2.

$$\text{probability of an event happening } = \frac{e^{(\alpha+\beta x)}}{1+e^{(\alpha+\beta x)}}, \tag{2}$$

where e is the natural logarithm base, α and β are parameters of the linear component of the model, and x is the value of the explanatory variable.

The following discussion deals with the calculation and interpretation of confidence intervals for fitted Y for both simple and multiple logistic regression models. The technique is described in detail with an emphasis on demonstrating how confidence intervals for fitted Y may be calculated in SPSS and interpreted. Information regarding the calculation and interpretation of confidence intervals for β are provided for completeness.

SIMPLE LOGISTIC REGRESSION

Confidence Intervals for β

Confidence intervals for β are provided by most statistical software packages, for reasonably large samples, using Equation 3.

$$\text{confidence intervals for } \beta = \hat{\beta} \pm 1.96(\text{ASE}), \tag{3}$$

where $\hat{\beta}$ is the estimated value of β, 1.96 is the large sample approximation of t for a two-tailed 95% confidence interval, and ASE is the asymptotic standard error of $\hat{\beta}$.

For logistic regression, β refers to the change in logit(p) that results from a unit change in x. Similarly, the confidence intervals for β show the upper and lower limits of the expected change in logit(p) for a unit change in x, assuming that the data are not under- or overdispersed. Logistic regression assumes errors with a binomial distribution, where the variance is a fixed function of the mean, var(y) = $\mu(1 - \mu)$, where μ is the mean of the response variable. This property of the variance being a function of the mean allows us to check to see if the distributional assumptions of the model are correct. Variance greater than would be expected for the assumed distribution is termed *overdispersion* and often arises because an important explanatory variable has not been measured, or because of dependencies induced by clustering in the process by which the data were sampled (e.g., students in schools, patients in treatment centers). Variance that is less than would be expected for the distribution is termed *underdispersion* and occurs more rarely, usually when there is some repulsion occurring in the data generating mechanism (e.g., territorial behavior leading to minimum distances between individuals or groups). A simple correction for over- or underdispersion involves adjusting the standard errors from the model using Equation 4.

$$\text{adjusted standard error} = s.e. \times \sqrt{(D/df)}, \tag{4}$$

where D is the deviance and df is the number of degrees of freedom of the model. This adjustment allows more realistic estimates of confidence intervals for the fitted model to be obtained but should be used with caution if the model is misspecified through the omission of an important explanatory variable—ideally one should try and find the source of the extra variation in the data and allow for this with additional terms in the model. Further details of modeling under- and overdispersed categorical data are given in Collett (1991) and Lindsey (1999).

Confidence Intervals for Fitted y

In addition to determining confidence intervals for the explanatory variable, confidence intervals for fitted values of y can also be obtained. These intervals are calculated for the linear predictor logit(p) (see Equation 1), and can be transformed to show confidence intervals for p (see Equation 2). The large sample approximation of the 95% two-tailed confidence intervals for the predicted mean value of logit(p) can be calculated using Equation 2 and are shown graphically in Figure 1. It is obvious from the graph that these intervals are not parallel but curved. This is because fitted y can be more accurately predicted when x is close to its mean value.

The confidence intervals for the probability of an event happening (p) can easily be obtained once the intervals for logit(p) have been derived. One simply transforms the graph so that the y axis represents p and not logit(p) (see Figure 2).

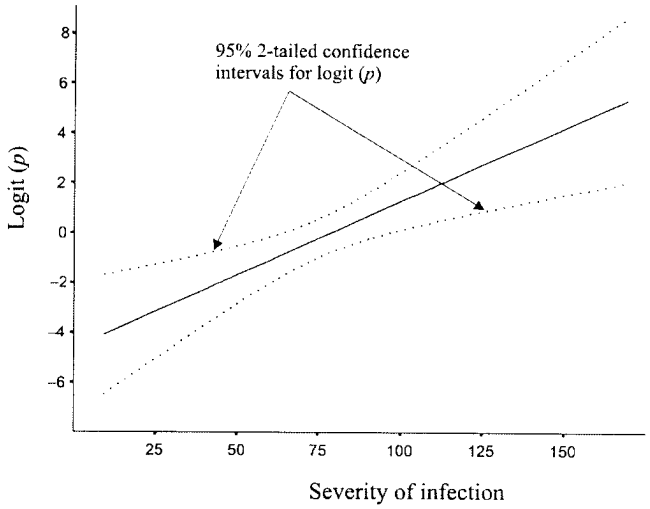

FIGURE 1 Confidence intervals for logit(p).

$$\text{confidence intervals for logit}(p) = \text{logit}(\hat{p}) \pm 1.96(\text{ASE}), \tag{5}$$

where logit($\hat{p}$) the estimated value of logit(p), ASE is the asymptotic standard error of logit($\hat{p}$) and 1.96 is the large sample approximation of t for a two-tailed 95% confidence interval.

Not all statistical packages provide confidence intervals for logit(p) directly, in which case they have to be calculated by hand. The difficulty with calculating these intervals is in determining the asymptotic standard error (ASE) associated with the prediction. We provide two techniques for calculating the ASE for logit(p), the first of which relies on the generation of a variance–covariance matrix, whereas the second is based on transforming the explanatory variable and recalculating the model. Both techniques are provided here because a variance–covariance matrix is not always provided, and in cases where it is, the calculation of the ASE can be complex, particularly for problems with multiple variables. The two methods also provide a means of comparison and validation.

Computing the ASE for logit(p) using a variance–covariance matrix. When the variance–covariance matrix is available, the ASE for logit(p) for a logistic regression model with a single explanatory variable can be calculated using Equation 6 (see Agresti, 1996, p. 110).

$$\text{large sample ASE for logit}(p) = \sqrt{\text{Var}(\hat{\alpha}) + x^2\text{Var}(\hat{\beta}) + 2x\text{Cov}(\hat{\alpha},\hat{\beta})}, \tag{6}$$

where $\mathrm{Var}(\hat{\alpha})$, $\mathrm{Var}(\hat{\beta})$, and $\mathrm{Cov}(\hat{\alpha}, \hat{\beta})$ are parameters obtained from the variance–covariance matrix, and x is the value of the explanatory variable. Values from the variance–covariance matrix and the value of the explanatory variable can be entered into Equation 6 and the ASE for the model calculated. Once this has been achieved, the confidence intervals for logit(p) can be determined using Equation 5. An example of the use of this method is provided in the following worked example.

Computing the ASE for Logit(p) using model refitting. The ASE for logit(p) can be calculated using a procedure in which one transforms the explanatory variable and refits the model (Crawley, 1993). Using this method, one can obtain an estimate of the ASE for any particular value of x. For example, the procedure for accomplishing this when $x = 50$ is as follows:

1. Subtract the value of the explanatory variable (x) you wish to use from each case of the original variable. For our example, each value of the explanatory variable should now equal $x - 50$.
2. Calculate a logistic regression model using the new values of the explanatory variable.
3. In the output statistics, the ASE associated with the constant (α) provides an estimate of the ASE for logit(p).

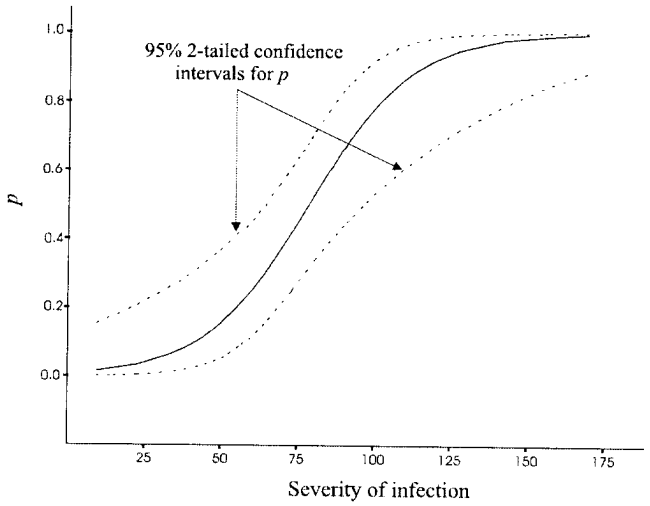

FIGURE 2 Confidence intervals for logit (p).

The ASE associated with logit(p) changes as a function of x and has to be calculated for each value of x we are interested in. An example of the use of this method to obtain the 95% two-tailed confidence intervals for logit(p) is provided below along with software commands for use with SPSS for Windows (Version 10.1).

A Worked Example of Simple Logistic Regression

We present here an example based on hypothetical data for illustration purposes. Treatment outcome (the response variable y) is modeled as a function of severity of infection (the explanatory variable x). For this demonstration we assume that the data are in an appropriate format and have been properly screened. The data file, of which two Versions are available labeled tab4_01.por (an SPSS portable data file) and tab4_01.dat (an ascii data file), can be downloaded as part of a zip file (hutsof99.zip) available from the StatLib datasets archive at http://lib.stat.cmu.edu/datasets/. The regression parameters for the model predicting treatment outcome from severity of infection (see Equation 7) are given in Table 1.

$$\text{logit}(p) = \alpha + \beta x = -4.640 + 0.059x \tag{7}$$

A unit increase in the severity of the infection (x) results in an increase of 0.059 in the log odds of the probability, logit(p), of the patient succumbing to the infection. Using the odds ratio, e^{β}, one can also state that for each unit increase in the severity of infection, the odds of succumbing to the disease increase by 6%. The large sample 95% two-tailed confidence intervals associated with β are 0.025 to 0.094, which means that in 95% of samples we would expect logit(p) to increase by somewhere between 0.025 and 0.094 for each unit increase in x. For each unit increase in infection severity, the odds of a patient succumbing to the infection are likely to increase somewhere between 3% and 10% ($e^{0.025}$ and $e^{0.094}$, respectively).

The confidence intervals for fitted y can also be calculated for this model once the ASE for logit(p) has been determined. Using the variance–covariance matrix in Table 2, the large sample ASE for logit(p) when $x = 50$ can be easily calculated.

TABLE 1
Model Parameters

			95% CIs for β		
	Coefficient	*ASE*	*Lower*	*Upper*	e^{β}
Severity	0.059	0.018	0.025	0.094	1.06
Constant	–4.640	1.383			

Note. CI = confidence interval; ASE = asymptotic standard error.

TABLE 2
Variance–Covariance Matrix

	Constant	*Severity*
Constant	1.89769	
Severity	–0.02326	0.00031

TABLE 3
Logistic Regression Model Parameters

	Coefficient	*ASE*	e^{β} *Odds Ratio*
$x - 50$	0.059	0.018	1.06
Constant	–1.680	0.583	

Note. ASE = asymptotic standard error.

Using model refitting, the ASE of logit(p) can be estimated by calculating the regression model using values of $x - 50$ instead of the original explanatory variable (x). The recalculated model is shown in Table 3. The estimate of the ASE for logit(p) is the ASE associated with the constant, which in this case is 0.583. This compares to the value of 0.589 calculated using the variance–covariance matrix.

Having calculated the ASE for logit(p) it is relatively simple to calculate the confidence intervals for fitted y. For this example, when $x = 50$, logit(p) is equivalent to $-4.64 + (0.059 \times 50)$, which equals -1.69. The large sample 95% two-tailed confidence intervals associated with this value of logit(p) can be determined using Equation 5. Using the ASE estimate derived from Table 2, the confidence intervals for logit(p) are the following:

$$\begin{aligned}\text{confidence intervals for logit}(p) &= \text{logit}(p) \pm 1.96(\text{ASE}) \\ &= -1.69 \pm (1.96 \times 0.583) \\ &= -2.83, -0.55.\end{aligned}$$

To understand more fully the predictions made from the model, all values can be transformed into direct measures of probability (instead of the log odds) by using the equation $p = e^{(\alpha + \beta x)}/(1 + e^{(\alpha + \beta x)})$. The probability of dying due to an infection of severity 50 is predicted to be 0.16 (see Equation 8).

$$\begin{aligned}\text{probability of death} &= \frac{e^{(\alpha+\beta x)}}{1+e^{(\alpha+\beta x)}} \\ &= \frac{e^{(-4.640+(0.059\times 50))}}{1+e^{(-4.640+(0.059\times 50))}} \\ &= 0.16.\end{aligned} \tag{8}$$

TABLE 4
Predictions Derived From the Regression Model

Infection Severity	*ASE for logit(p)*	*logit(p)*			*Probability of Dying*		
		Predicted	*95% CIs*		*Predicted*	*95% CIs*	
Using variance–covariance matrix							
0	1.377	–4.64	–7.34	–1.94	0.01	0.00	0.13
50	0.589	–1.69	–2.84	–0.54	0.16	0.06	0.37
100	0.588	1.26	0.11	2.41	0.78	0.53	0.92
150	1.376	4.21	1.51	6.91	0.99	0.82	1.00
200	2.235	7.16	2.78	11.54	1.00	0.94	1.00
Using model refitting							
0	1.383	–4.64	–7.35	–1.93	0.01	0.00	0.13
50	0.583	–1.69	–2.83	–0.55	0.16	0.06	0.37
100	0.558	1.26	0.17	2.35	0.78	0.53	0.92
150	1.352	4.21	1.56	6.86	0.99	0.83	1.00
200	2.211	7.16	2.83	11.45	1.00	0.94	1.00

Note. ASE = asymptotic standard error; CI = confidence interval.

It is an easy matter to recalculate the ASE for different values of x and obtain predictions and confidence intervals for fitted y for a whole range of values of the explanatory variable. Table 4 shows the ASEs, the logits, and the probabilities associated with the predicted values of the response variable and the 95% two-tailed confidence intervals for a number of different infection severities using ASE values calculated from both methods demonstrated previously. It is interesting to note how similar the results are for the two methods (in fact, the predicted probability of dying is identical to two decimal places).

MULTIPLE LOGISTIC REGRESSION

Confidence Intervals for β

As is the case for simple logistic regression, confidence intervals for β can be computed using Equation 3 with the value of 1.96 being suitable for samples of 30 or more cases (Agresti & Finley, 1997). With smaller samples it is recommended that one use Student's t distribution to calculate confidence intervals (e.g., Crawley, 1993). For multiple logistic regression, β refers to the change in logit(p) that results from a unit change in x when all other variables in the model are held constant. Similarly, the confidence intervals for β show the upper and lower limits of the expected change in logit(p) for a unit change in x, when all other variables are held constant, assuming that the data are not under- or overdispersed.

Confidence Intervals for Fitted *y*

Confidence intervals for models containing multiple explanatory variables can be calculated for fitted *y* in much the same way as they were in the case in which there was only one explanatory variable. These intervals are first calculated for logit(p) (see Equation 5) and are then transformed to represent the intervals for p. As in the case of simple logistic regression we provide two techniques for computing the ASE for logit(p).

Computing the ASE for logit(p) using a variance–covariance matrix. The large sample ASE of logit(p) may be calculated using Equation 9 (see Collett, 1991, p. 88). As can be seen, the calculation is complex. Once the ASE for logit(p) has been determined for a particular combination of x values, it can be entered into Equation 5 and the confidence intervals can be determined. An example of the use of this method is provided in the following worked example.

$$\text{ASE of logit}(p) = \sqrt{\sum_{j=1}^{k} x_{j0}^2 \text{Var}(\hat{\beta}_j) + 2\sum_{j=1}^{k}\sum_{h=1}^{j} x_{h0}x_{j0}\text{Cov}(\hat{\beta}_h, \hat{\beta}_j)}, \qquad (9)$$

where $\text{Var}(\hat{\beta}_j)$ and $\text{Cov}(\hat{\beta}_h, \hat{\beta}_j)$ are parameters obtained from the variance–covariance matrix, and x is the value of the explanatory variable.

Computing the ASE for logit(p) using model refitting. The ASE for logit(p) can also be calculated using a procedure that recalculates the regression model using transformed values of each explanatory variable. For a multiple logistic regression, the procedure for accomplishing this is as follows:

1. Create new explanatory variables by subtracting the chosen value of the explanatory variable from each case of the original variable.
2. Recalculate the model using the new explanatory variables.
3. The ASE associated with the constant provides an estimate of the ASE for logit(p).

An example of the use of this method to obtain the 95% two-tailed confidence intervals for logit(p) is provided next along with software commands for use with SPSS for Windows (Version 10.1).

A Worked Example of Multiple Logistic Regression

The primary purpose of this example is to demonstrate the logistic regression procedure rather than provide a best model. The data to be used was taken from a study

on child eyewitness testimony (Hutcheson, Baxter, Telfer, & Warden, 1995) and modified so that certain procedures can be demonstrated. The data file, of which two versions are available labeled logis_d.por (an SPSS portable data file) and logis_d.dat (an ascii data file), can be downloaded as part of a zip file (hutsof99.zip) available from the StatLib datasets archive at http://lib.stat.cmu.edu/datasets/.

The overall aim is to model the probability that a case will be heard in court (the dichotomous variable, prosecute) given certain information about the child and details about the child's testimony. The explanatory variables that are included in this model are the child's age (children from two different age groups took part in the study: 5- to 6-year-old children attending Primary 1 and 8- to 9-year-old children attending Primary 4), the overall quality of the testimony (based on its completeness and accuracy), and the location of the interview (the child's home, school, in a standard police interview room, or a specially constructed children's interview room). Location was dummy-coded using the indicator coding method with the standard police interview room as the reference category. The factor age has also been dummy coded using the indicator method so that the coefficients generated by software refer to comparisons between age groups (8- to 9-year-olds are the reference category).

The overall model fit as shown in Table 5 shows that the explanatory variables together contribute significantly to the prediction of the response variable. It should be noted that as there are a number of variables likely to be important for prosecution that have not been included in the analysis (e.g., type of crime, level of violence, and the child's fear of repercussion if evidence is given), the fit of the model might not appear particularly good, indicated by the model chi-square statistic and the associated p value. However, the use of a simpler model in this instance can be justified for the purpose of a clear demonstration of the methods of calculating confidence intervals for the fitted values.

TABLE 5
Predicting Prosecution

				95% CIs for e^{β}	
	Coefficient	*ASE*	e^{β}	*Lower*	*Upper*
Quality	0.050	0.033	1.051	0.984	1.122
Age	–0.907	0.637	0.404	0.116	1.409
Location					
Home	–0.452	0.752	0.636	0.150	2.775
School	0.070	0.758	1.072	0.243	4.739
Special	0.773	0.810	2.166	0.442	10.601
Constant	–2.820	2.168			

Note. ASE = asymptotic standard error; CI = confidence interval; logit(p) = –2.820 + 0.05quality – 0.907age – 0.452home + 0.07school + 0.773special. Model chi- square = 15.729, $df = 5$ $p = .0077$.

TABLE 6
Variance–Covariance Matrix

	Constant	*Age*	*School*	*Home*	*Special*	*Quality*
Constant	4.667920					
Age	–0.817716	0.404895				
School	–0.178822	0.021321	0.572003			
Home	–0.135454	0.003707	0.283277	0.561751		
Special	0.231957	–0.065487	0.291245	0.295162	0.653605	
Quality	–0.068954	0.010732	–0.001846	–0.002448	–0.008209	0.001100

The regression parameters shown in Table 5 illustrate the effects that each of the variables has in determining whether a case is prosecuted. For example, compared to 8- to 9-year-old children, testimonies from 5- to 6-year-olds are only 0.404 as likely (the odds ratio e^{β}) to lead to prosecution. This result is nonsignificant as the 95% confidence intervals for e^{β} includes 1.0 (no change). Predictions can be made from the model about the likelihood of a case being heard in court. For example, the probability of a testimony of a 5- to 6-year-old child, interviewed at home and providing a testimony with a quality of 50, being heard in court is predicted to be as follows:

$$\begin{aligned}\text{logit}(p) &= \alpha+\beta_1\text{quality}+\beta_2\text{age}+\beta_3\text{home}\\ &= -2.82+(0.05\times 50)-0.907-0.452\\ &= -1.679.\\ p &= e^{-1.679}/(1+e^{-1.679}) = 0.157.\end{aligned}$$

Estimated values of p are, however, considerably more meaningful when quoted in conjunction with their confidence intervals. Two methods for estimating them are demonstrated here, one using the variance–covariance matrix for the model and one refitting the model using transformed explanatory variables.

The variance–covariance matrix for the previous model is shown in Table 6. The large sample ASE for logit(p) can be calculated using Equation 9 and our designated values for age (1), location (home = 1, all others 0), and quality (50):

$$\begin{aligned}\text{ASE} &= \sqrt{4.668+0.405+0.562+(50^2\times 0.001)+2(-0.818-0.135+(50\times -0.069)}\\ &\quad \sqrt{+0.004+(50\times 0.011)+(50\times -0.002))}\\ &= 0.647.\end{aligned}$$

Using the original values from the table (for the purpose of reducing rounding error), the ASE for the model equals 0.647. Inputting this value into Equation 5 yields

TABLE 7
Estimating the ASE for logit(p)

	Coefficient	*ASE*	e^{β}	*95% CIs for* e^{β}	
				Lower	*Upper*
Quality – 50	0.050	0.033	1.051	0.984	1.122
Age (0 to 1)	–0.907	0.637	0.404	0.116	1.409
Location					
Home – 1	–0.452	0.752	0.636	0.150	2.775
School	0.070	0.758	1.072	0.243	4.739
Special	0.773	0.810	2.166	0.442	10.601
Constant	–1.705	0.650			

Note. ASE = asymptotic standard error; CI = confidence interval. Model chi-square = 15.729, df = 5, p = .0077.

confidence intervals for logit(p) of $-1.679 \pm (1.96 \times 0.647)$, giving –2.947, –0.411. Converting these into estimates for p ($e^{-2.947}/(1 + e^{-2.947})$ and $e^{-0.411}/(1 + e^{-0.411})$ provides intervals of 0.050, 0.399).

Confidence intervals for logit(p) can also be determined by recalculating the model using transformed variables. Predicting the probability of a case being heard in court for a 5- to 6-year-old child, interviewed at home and who obtains a score of 50 for quality, one recodes[1] age from 0, 1 to 0, –1, deletes 1 from the dummy variable home (the reference category remains the same), and deletes 50 from quality. These new variables are then entered into the model in place of the original variables. This model is shown in Table 7. The ASE for the constant term provides the estimate of the ASE for logit(p). For this example, the ASE is predicted as 0.650 (compared to 0.647 using the variance–covariance matrix). The 95% confidence intervals for logit(p) are therefore $-1.679 \pm (1.96 \times 0.650)$, giving intervals of –2.953 and –0.405 for logit(p), with the corresponding confidence intervals for p being equivalent to 0.050, 0.400.

CONCLUSIONS

In this article we have demonstrated a method to calculate confidence intervals for predicted values of the response variable in logistic regression models. The method utilizing data transformation and model refitting is particularly useful as it can be applied in statistical packages that do not compute variance–covariance matrices or confidence intervals directly (SPSS, being a notable example). The technique was

[1]The reference category remains the same, but the coding is inverted (i.e., 1, 0 becomes –1, 0) providing the correct sign for the Age parameter.

validated using a procedure that calculated the intervals directly from a variance–covariance matrix. The lack of automated computation of confidence intervals for predictions from logistic regression models (and the related multinomial logistic regression and proportional odds models) in some popular statistical packages is a serious omission. This omission is compounded by the almost exclusive reliance on SPSS for teaching and analysis in a number of social science fields. Until these intervals are provided directly by software, researchers will need to compute them manually using the data transformation and model-recalculation technique described in this article.

REFERENCES

Agresti, A. (1990). *Categorical data analysis.* Chichester, England: Wiley.

Agresti, A. (1996). *An introduction to categorical data analysis.* Chicester, England: Wiley.

Agresti, A., & Finley, B. (1997). *Statistical methods for the social sciences* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Collett, D. (1991). *Modelling binary data.* London: Chapman & Hall.

Crawley, M. J. (1993). *GLIM for ecologists.* Oxford, England: Blackwell Science.

Hutcheson, G. D., Baxter, J. S., Telfer, K., & Warden, D. (1995). Child witness statement quality: General questions and errors of omission. *Law and Human Behaviour, 19,* 631–648.

Lindsey, J. K. (1999). *Models for repeated measurements* (2nd ed.). Oxford, England: Oxford University Press.

SPSS. (1999a). *SPSS base 10.0 user's guide.* Chicago: Author.

SPSS. (1999b). *SPSS regression models 10.0.* Chicago: Author.

APPENDIX
Statistical Software Commands

Next are the SPSS software commands (Version 10.1) required to run the analyses described in this article. Further details of the SPSS statistical package can be found in the appropriate manuals (see, e.g, SPSS, chap. 5 and 6; 1999b, chap. 2 and 8).

Simple Logistic Regression Example

Load the data file tab4_01.por into SPSS. This file contains two variables, severity (a continuous variable denoting infection severity) and outcome (a binary classification of treatment outcome).

Calculating the logistic regression model.

Analyze ▼
Regression ▶

Binary Logistic ...
Dependent: *input* Outcome
Covariates: *input* Severity
Save ...
Predicted Values: Probabilities: *check box*
Continue
OK

The previous commands will compute the regression model and save predicted values for p under the variable name pre_1. The regression equation previously calculated is as follows:

$$\text{logit}(p) = -4.6401 + 0.0592\text{Severity}.$$

Calculating Confidence Intervals for Fitted y

To calculate the confidence intervals for fitted y we first need to determine the ASE associated with logit(p). In SPSS, when severity is equal to 50, this can be calculated using the model refitting method:

First, compute a new variable equal to severity – 50 (Call it sev_50) and use this variable in the following model:

Analyze ▼
Regression ▶
Binary Logistic ...
Dependent: *input* Outcome
Covariates: *input* Sev_50
Continue
OK

The ASE for logit(p) is given by the ASE for the constant, which in this case is 0.583. Using this value, one can calculate the confidence intervals for logit(p) and p (see the explanation provided in the text).

Multiple Logistic Regression Example

Load the data file logis_d.por (an SPSS portable data file) into SPSS. This file contains a number of variables, only a few of which are to be used in this analysis. The

variables of interest are prosecute (a binary classification of whether a case gets to court), age (a binary classification of two age groups), and location (an unordered categorical variable indicating four different locations). The variable location has been represented as four dummy variables, loc_home, loc_scho, loc_form, and loc_spec, to indicate the four different locations used: home, school, formal interview room, and a specially designed interview room. Although there are four dummy variables provided, only three can be entered into the regression model at one time.

The codes used are as follows:

Prosecute: no = 0, yes = 1.

Age: 5- to 6-year-old children = 0, 8- to 9-year-old children = 1.

Location:
loc_home: home = 1, not home = 0
loc_scho: school = 1, not school = 0
loc_form: formal interview room = 1, not formal interview room = 0
loc_spec: special interview room = 1, not special interview room =0

Calculating the logistic regression model.

Run the logistic regression model using the following commands:
Analyze ▼
Regression ▶
Binary Logistic ...
Dependent: *input* Prosecute
Covariates: *input* Age, Quality, loc_home, loc_scho *and* loc_spec
Categorical ...
Categorical Covariate: *input* Age
Continue
OK

which should give the same parameters as shown in Table 5.

Calculating Confidence Intervals for Fitted y

To calculate the ASE when there are multiple variables, the model refitting procedure can be used. Using the example in A Worked Example of Multiple Logistic Regression section, the ASE for logit(p) can be calculated for a 5- to 6-year-old child, interviewed at home and producing an interview of quality 50 by creating

new variables for age, home, and quality. Age is recoded from 0, 1 to 0, –1, loc_home is recoded to (loc_home – 1) and quality to (Quality –50). Call the variables Qual_new, Home_new, and Age_new. Input these variables into the logistic regression model:

Statistics ▼
Regression ▶
Logistic ...
Dependent: *input* Prosecute
Covariates: *input* Age_new, Qual_new, home_new, loc_scho, and loc_spec
OK

The ASE for logit(p) is given by the ASE for the constant, which in this case is 0.650. Using this value, one can calculate the confidence intervals for logit(p) and p (see the explanation provided in the text).