Journal of Official Statistics
Vol. 3, No. 4, 1987, pp. 419–429
© Statistics Sweden

Correction for Misclassification Using Doubly Sampled Data

Anders Ekholm[1] *and Juni Palmgren*[2]

Abstract: In doubly sampled data the units of a subsample are classified jointly by two methods: (i) a fallible but inexpensive, and (ii) a reliable but expensive. The rest of the units are classified only by method (i). We propose an extension of the generalized linear model (Nelder and Wedderburn (1972)) for such data. We model explicitly the nonsampling errors, i.e., the probabilities of misclassification. We then incorporate these into the model for the dependence of the response on the explanatory factors. There might be misclassifications both in the response and in the explanatory factors.

A car accident data set is analyzed in which 80 084 accidents were categorized only by the police, and 1 796 accidents were categorized both by the police and by personal interview of the accident victims. Our model is more explicit concerning the nonsampling errors than the models used for these data by Hochberg (1977) and by Espeland and Odoroff (1985).

Key words: Error in explanatory factor; error in binary response; exponential family nonlinear model; generalized linear model; GLIM; misclassification model; structural model.

1. Introduction

We start with the fictitious example to convey the idea of doubly sampled data and of our method of analyzing it. The task is to estimate the probability of a certain back disorder in middle-aged men. There are two ways to diagnose this disease; by a very time consuming medical examination or by using a questionnaire. The former method is fully reliable but expensive, the latter is handy but error prone. We have data from 1 000 men of which a random subsample of 100 were subjected both to the medical examination and to the questionnaire, while the remaining 900 were required to answer the questionnaire only. (See Table 1.)

There are at least two naive ways of estimating the probability of the disorder. Using only the examination data (D), the estimate is 0.3, with a standard error (s.e.) of 0.046. On the other hand, using only the questionnaire (D^*), we compute the estimate as $(33 + 298)/(100 + 900) = 0.331$ with an s.e. of 0.015. The former estimate is free from bias due to nonsampling error but suffers from a large standard error. The latter estimate is affected by nonsampling error but has a much smaller standard error. Neither method uses information from the cross-classification of D and D^*.

Our approach is to write down a saturated model for the left hand 2 × 2 table in Table 1, and to derive the probabilities of the right

[1] Associate Professor of Statistics at the University of Helsinki, Finland.
[2] Biostatistician at the National Public Health Institute, Finland.

Acknowledgement: Financial support from the Academy of Finland is gratefully acknowledged.

Table 1. Fictitious data concerning the occurence of back disorder among middle-aged men. D refers to the medical examination, and D to the questionnaire result. D– and D+ stand for negative and positive examination result respectively, and analogously for D**

	Both examined and questioned			Only questioned
	$D+$	$D-$	Total	
D^*+	24	9	33	298
D^*-	6	61	67	602
Total	30	70	**100**	**900**

Table 2. A model for the probabilities of the cells of Table 1

	Both examined and questioned		Only questioned
	$D+$	$D-$	
D^*+	$\pi(1-\delta)$	$(1-\pi)\varepsilon$	$\pi(1-\delta)+(1-\pi)\varepsilon$
D^*-	$\pi\delta$	$(1-\pi)(1-\varepsilon)$	$\pi\delta+(1-\pi)(1-\varepsilon)$
Sum	π	$1-\pi$	1

hand collapsed 2 × 1 table from this. Let π represent the true probability of disorder, and δ and ε the probabilities of false negative and false positive diagnosis respectively. That is

$$\pi = \mathrm{pr}(D+), \quad \delta = \mathrm{pr}(D^*-\,|\,D+),$$
$$\varepsilon = \mathrm{pr}(D^*+\,|\,D-).$$

The probabilities for the cells of Table 1 are shown in Table 2.

Table 1 contains six frequencies restricted by the fixed sums 100 and 900. There are thus four independent observations. Table 2 specifies a model with three parameters for these observations. We estimate the three parameters by maximum likelihood, and arrive at the following estimates for π, δ, and ε respectively, with the standard errors in parenthesis 0.3 (0.036), 0.2 (0.065), and 0.13 (0.033).

Note that we obtain maximum likelihood estimates of the structural parameter π and of the error probabilities, all based on the full data set. In this simple example there is a single structural parameter. In more realistic cases we might want to fit a structural model to describe the effect of explanatory factors on a categorical response. The presence of doubly classified units makes it possible to model the nonsampling errors jointly with the structural model. We refer to the restrictions imposed on the probabilities of errors as the misclassification model. The complete model description is the combination of the structural model and the misclassification model. Large sample goodness-of-fit tests are available for testing the full model.

In Section 2 we show that the full model can be regarded as an extension of the generalized linear models. The estimation is, in fact, conveniently performed in GLIM (Payne (1985)). In Section 3 we present a large doubly sampled data set from highway safety research. In Section 4 we analyze these data using the type of model described above. In Section 5 we briefly discuss the suitability of double sampling for making structural inference from data sets originally collected for administrative purposes.

2. Exponential Family Nonlinear Models

We shall give a condensed description of generalized linear models as applied to multidimensional contingency tables. McCullagh and Nelder (1983, Ch. 6) give a sophisticated presentation, and Dobson (1983, Ch. 9) gives an instructive introduction.

Let $Y_1, \ldots ,Y_n$ denote the random counts in the n cells of a multidimensional contingency table. Their joint distribution is multinomial if the grand total of the Y_is is fixed, and product multinomial if several subtotals are also fixed. We treat subtotals as fixed, either because the data are collected keeping these subtotals fixed, or because we want to condition the analysis on observed counts in the margin formed by cross-classifying explanatory factors.

In both the multinomial and the product multinomial cases we treat, as a technical simplification, the Y_is as independent Poisson variates. Palmgren (1981) shows that correct large-sample maximum likelihood inference about any structure on the probabilities in the multinomial or product multinomial distribution is obtained from the Poisson assumption as long as the following condition is fulfilled. The sum of the fitted counts for any fixed margin must coincide with the corresponding observed sum. The random variation of the cell counts is then adequately described by the Poisson distribution.

A generalized linear model for the systematic variation of the cell probabilities expresses $\ln(E(Y_i))$ for $i = 1, \ldots , n$ as a sum of parameters. If, for example, the contingency table is $r \times c$, and we want to model independence between the classifying factors, then we write

$$\ln(E(Y_i)) = \alpha + \beta_u + \gamma_v,\ i = 1,\ldots,n; \\ u = 1,\ldots, r\ ;\ v = 1,\ldots, c,$$

where α denotes a parameter common for all cells, and the β_us and γ_vs are sets of row and column main effects. Restrictions must be set on the parameters to make the model identifiable. For tables with many dimensions and several possible layers of interaction terms, more parameters are included, but always by simple addition.

In the terminology of the generalized linear model, the transformation ln() applied to the expectation of the Y_is is called the link function. The logarithm might be substituted by some other monotonic and differentiable function, but the link of the expectation is always expressed as a linear function of the parameters.

Now, consider Tables 1 and 2. There we have a set of random counts modeled so that their fitted values will coincide with the fixed totals 100 and 900. The snag is that their expectations are expressed as products of the parameters (in the left hand table) and sums of products (in the right hand table). Accordingly, neither the logarithm nor any other regular function applied to the expectations will give us an expression which is a linear function of the parameters.

Palmgren and Ekholm (1987) introduced a class of models called exponential family nonlinear models. The random variation is modeled in exactly the same way as for generalized linear models. The model for the systematic variation is, however, more general in two ways: (i) the link function is defined separately for each cell of the table under study, (ii) there is no requirement for any expression which is a linear function of the parameters. Ekholm et al. (1986) have written a set of macros for fitting exponential family nonlinear models in GLIM. The user just has to program the formula for his model in the GLIM language. GLIM then finds maximum likelihood estimates of the parameters, and offers all its standard provisions. We illustrate the usefulness of these for statistical analysis in Section 4.

3. The Seat Belt Data

To illustrate exponential family nonlinear modeling of a structural relationship with superimposed misclassification, we reanalyze a data set first presented and analyzed by Hochberg (1977). He does not consider any restrictions for the misclassification probabilities. Espeland and Odoroff (1985) reanalyzed these data by a two stage EM-procedure. At the first stage they identify a misclassification model assuming a saturated structural model. At the second stage they identify a nonsaturated structural model using the misclassification model they found at the first stage. Chen et al. (1984) added an artificial data set to the original one, and then analyzed it as triply sampled data. We restrict our attention to the doubly sampled real data. There are two new features of our analysis. (i) We formulate and estimate one single model, which includes explicit parameters and restrictions for the error probabilities. (ii) We use a quickly converging Newton-Raphson algorithm for maximum likelihood estimation providing standard errors and correlations for estimates of all the parameters.

The singly classified sample consists of 80 084 accidents that were fully recorded by the police in North Carolina during 1974. The police records classify the accidents according to the following four factors:

A^* = Driver's seat belt usage; A^*- for no,
A^*+ for yes,
B^* = Driver's injury; B^*- for no,
B^*+ for yes,
C = Driver's sex; $C-$ for male and
$C+$ for female,
D = Damage; $D-$ for low, $D+$ for high.

The doubly classified sample consists of 1 796 accidents recorded in the beginning of 1975 in North Carolina. These accidents are classified by the police according to the four factors above, and also according to two additional factors, which we shall refer to as:

A = true seat belt usage; $A-$ for no,
$A+$ for yes,
B = true injury; $B-$ for no, $B+$ for yes.

These "true" classifications are (Hochberg (1977, p. 919)) based on intensive inquiries for each individual case. The reliable classifiers

Table 3. Number of road accidents in North Carolina classified according to year, driver's sex (C), seat belt usage (A and A), injury (B* and B), and severity of damage (D). Seat belt usage and injury are classified by police and by inquiries respectively. Fitted values from final model in parentheses*

Layer 1: $C-$, $D-$, (male, low damage)

		1975				1974
		$A-$		$A+$		
		$B-$	$B+$	$B-$	$B+$	
A^*-	B^*-	407 (408.0)	45 (44.3)	62 (58.3)	7 (4.5)	22 536 (22 530.7)
	B^*+	5 (5.6)	32 (29.7)	1 (0.8)	4 (3.0)	1 687 (1 708.7)
A^*+	B^*-	6 (7.9)	1 (0.9)	47 (55.6)	6 (4.3)	3 006 (3 000.5)
	B^*+	0 (0.1)	1 (0.6)	1 (0.8)	2 (2.8)	199 (188,1)
		497		**130**		**27 428**

Layer 2: $C-$, $D+$, (male, high damage)

		1975				1974
		$A-$		$A+$		
		$B-$	$B+$	$B-$	$B+$	
A^*-	B^*-	299 (276.3)	59 (71.8)	20 (20.9)	9 (3.8)	17 476 (17 448.8)
	B^*+	11 (13.8)	118 (121.7)	2 (1.0)	5 (6.5)	6 746 (6 695.8)
A^*+	B^*-	4 (5.4)	1 (1.4)	30 (32.9)	6 (6.0)	2 155 (2 137.0)
	B^*+	1 (0.3)	0 (2.4)	2 (1.6)	9 (10.2)	583 (678.5)
		493		**83**		**26 960**

Layer 3: $C+$, $D-$, (female, low damage)

		1975				1974
		$A-$		$A+$		
		$B-$	$B+$	$B-$	$B+$	
A^*-	B^*-	206 (201.9)	37 (42.5)	18 (14.1)	5 (2.1)	11 199 (11 329.5)
	B^*+	4 (2.8)	29 (28.5)	0 (0.2)	0 (1.4)	1 422 (1 428.2)
A^*+	B^*-	1 (3.9)	0 (0.8)	17 (18.4)	1 (2.7)	1 262 (1 125.3)
	B^*+	3 (0.05)	1 (0.6)	0 (0.2)	0 (1.8)	117 (117.1)
		281		**41**		**14 000**

Layer 4: $C+$, $D+$, (female, high damage)

		1975				1974
		$A-$		$A+$		
		$B-$	$B+$	$B-$	$B+$	
A^*-	B^*-	102 (109.4)	53 (45.8)	7 (5.0)	4 (1.5)	6 964 (6 977.7)
	B^*+	5 (5.5)	79 (77.7)	1 (0.2)	1 (2.5)	3 707 (3 707.5)
A^*+	B^*-	2 (2.1)	1 (0.9)	6 (10.2)	3 (3.0)	728 (699.0)
	B^*+	0 (0.1)	1 (1.5)	0 (0.5)	6 (5.1)	297 (311.8)
		243		**28**		**11 696**

are denoted by a single letter, while the error prone classifiers bear a star. The full data set is reported in Table 3. We use the same framework as in the introductory example; the number of doubly classified units in the left hand side, and the collapsed data in the right hand side. The bold numbers in Table 3 are the totals in the three dimensional (A, C, D) margin for the 1975 data and in the two dimensional (C, D) margin for the 1974 data. We will treat these totals as fixed numbers in the analysis.

The relevant structural question to ask about these data is whether the use of seat belt has any effect on the probability that the driver is injured, adjusting for the driver's sex and the degree of damage. Thus the factors A, C, and D are explanatory while B is a response. In a technical sense we work with a three dimensional $2 \times 2 \times 2$ response (A^*, B^*, B). In the substantive sense, however, B is the single response contaminated by misclassifications, and A, C, and D are explanatory factors. The explanatory factor A is contaminated too. Treating the (A, C, D) margin as fixed is equivalent to making inference conditionally on the observed outcomes of the explanatory factors, cf. Palmgren (1981).

First we neglect misclassification, and concentrate on the 1975 data, where we have access to the true classifications A and B. Table 4 summarizes the seat belt effect by reporting the relative frequencies of injury for a cross classification of sex, damage, and seat belt usage. These figures are obtained from the left hand parts of Table 3 by summing over factors A^* and B^*. The first entry in Table 4 is obtained as 79/497 = 0.16.

Table 4. The relative frequencies of injury for different cross classified levels of Sex (C), Damage (D), and Seat belt usage (A)

		$A-$	$A+$
$C-$	$D-$	0.16	0.15
	$D+$	0.36	0.35
$C+$	$D-$	0.24	0.15
	$D+$	0.55	0.50

The relative frequencies of injury are consistently lower for seat belt users, but this effect is not statistically significant. We have fitted a series of logit models to these data. Four models are reported in Table 5.

A test statistic for the seat belt effect is the difference in deviances between models 1 and 2, which is 1.08 on 1 degree of freedom and not significant. Note that the same conclusion is reached by comparing models 3 and 4. Further, the estimate for the A-effect is the same from models 1 and 4. If we stick to just the true classifications, we must conclude that the effect of using seat belts is negligible.

Table 5. Four logit models for the 1975 data marginalized over the police classifications of Seat belt usage (A^*) and Injury (B^*)

Model	Deviance	df	p	Estimate of A-effect	s.e.
1. $C+D+C.D+A$	1.24	3	0.74	–0.16	0.157
2. $C+D+C.D$	2.32	4	0.68		
3. $C+D$	4.10	5	0.54		
4. $C+D+A$	3.09	4	0.54	–0.16	0.157

Comparing models 2 and 3 we find that $C.D$, the interaction between C and D, is nonsignificant. No further simplification can be made without considerable misfit. The estimates of the sex (C) and the damage (D) main effects are (with standard errors in parentheses) respectively 0.63 (0.11) and 1.22 (0.11). The probability of injury is higher for female than for male drivers, and higher when the damage is high. The damage effect is easy to understand, but we do not know the reason for the sex effect.

4. Misclassification and the Seat Belt Effect

We turn to the full data, and build a misclassification model on top of a logit model for the probability of injury. To present this much more complicated model we introduce some new notation. Let $a = +$ or $-$ be a generic symbol for the levels of factor A, and use c and d analogously for the levels of factors C and D. Let π represent the structural probability of injury, that is,

$$\pi_{acd} = \mathrm{pr}(B+ \mid A=a,\ C=c,\ D=d).$$

Next, we define four different types of misclassification, which might be functions of the levels of factors C and D:

$$\begin{aligned}
\beta_1(c, d) &= \mathrm{pr}(B^*+ \mid B-, \quad C=c, \quad D=d),\\
\beta_2(c, d) &= \mathrm{pr}(B^*- \mid B+, \quad C=c, \quad D=d),\\
\alpha_1(c, d) &= \mathrm{pr}(A^*+ \mid A-, \quad C=c, \quad D=d),\\
\alpha_2(c, d) &= \mathrm{pr}(A^*- \mid A+, \quad C=c, \quad D=d).
\end{aligned}$$

The reader will find it useful to have available the following list of the four types of misclassification associated with the above probabilities:

β_1: Police report **injury**, when inquiry reports **no injury**,

β_2: Police report **no injury**, when inquiry reports **injury**,

α_1: Police report **belt**, when inquiry reports **no belt**,

α_2: Police report **no belt**, when inquiry reports **belt**.

In Table 6 we present the expressions for the cells of the left hand parts of Table 3. It will suffice to do so for Layer 1, that is for $C-$, $D-$. Since C and D are fixed, we suppress their indices, and the index for A is $+$ or $-$, as the case may be. For example, π_+ denotes the conditional probability of injury given that the seat belt was used.

The individual probabilities in Table 6 are derived by elementary rules of probability calculus just like the probabilities in the introductory example. Consider the entry in row A^*-, B^*+ and column $A+$, $B-$. The entry is $(1 - \pi_+)\,\beta_1\alpha_2$, which is the conditional probability given that the belt was used, that no injury

Table 6. The expressions for the probabilities of the cells on the left side of Table 3

		$A-$		$A+$	
		$B-$	$B+$	$B-$	$B+$
A^*-	B^*-	$(1-\pi_-)(1-\beta_1)(1-\alpha_1)$	$\pi_-\beta_2(1-\alpha_1)$	$(1-\pi_+)(1-\beta_1)\alpha_2$	$\pi_+\beta_2\alpha_2$
	B^*+	$(1-\pi_-)\beta_1(1-\alpha_1)$	$\pi_-(1-\beta_2)(1-\alpha_1)$	$(1-\pi_+)\beta_1\alpha_2$	$\pi_+(1-\beta_2)\alpha_2$
A^*+	B^*-	$(1-\pi_-)(1-\beta_1)\alpha_1$	$\pi_-\beta_2\alpha_1$	$(1-\pi_+)(1-\beta_1)(1-\alpha_2)$	$\pi_+\beta_2(1-\alpha_2)$
	B*+	$(1-\pi_-)\beta_1\alpha_1$	$\pi_-(1-\beta_2)\alpha_1$	$(1-\pi_+)\beta_1(1-\alpha_2)$	$\pi_+(1-\beta_2)(1-\alpha_2)$
Sum		$1-\pi_-$	π_-	$1-\pi_+$	π_+

occurred, but the police classified the accident as "no belt and injury." Note that the probabilities in the $A-$ half sum to 1, and so do the probabilities in the $A+$ half. This implies that the $A-$ totals and the $A+$ totals of the fitted frequencies will coincide with the observed frequencies for each given level of sex and damage. That is, the (A, C, D) marginal totals are fixed at their observed values.

Next we calculate the probabilities for the cells of the 1974 data, cf. Table 3. We sum the probabilities in Table 6 across rows, weighting the probabilities in the $A-$ half and in the $A+$ half by the respective proportions of the $A-$ and the $A+$ totals from the corresponding layer in Table 3. For Layer 1 the weight for the $A-$ half is 497/(497 + 130), and the weight for the $A+$ half is 130/627. This particular weighting is just a further consequence of treating the (A, C, D) margin as fixed. It follows that the probabilities in each layer of the 1974 table sum to 1, and the (C, D) totals of observed and fitted frequencies coincide. The physical meaning of the weighting we use is that we assume that the true proportion of belt users is the same in the 1974 and in the 1975 data. There is no obvious reason to doubt this assumption, if the 1975 accidents are a random subsample of all the accidents.

Table 6 shows that there are six parameters for each layer of Table 3. Having imposed a structure for the parameters across layers, the model is written in GLIM language, and the procedure of Ekholm et al. (1986) is applied. This is a straightforward exercise, and we have, in fact, estimated a considerable number of different models for these data. There might, as always, be other well-fitting models, but we have found no obvious contender.

The deviance of our best model is 58.91 on 52 degrees of freedom corresponding to a significance level of 0.24. The Pearson test statistic has the value 56.25. We report this model in Tables 7 and 8.

Table 7. The parameters of the structural logit model for the data in Table 3. The first level of each parameter is set to zero, and **1** *is the logit of the probability for A–, C–, D–*

Parameter	Estimate	s.e.
1	–1.721	0.0775
$C+$	0.6629	0.0571
$D+$	1.316	0.0940
$C+.D+$	–0.1854	0.0695
$A+$	–0.3480	0.0799

The structural model is still expressed in logit units in Table 7. The seat belt (A) effect is now more than double compared to the one in Table 5. It is clearly significant. No other interactions except the one between sex and damage $(C.D)$ are significantly different from zero. We comment further on the structural part below, but first we discuss the error probabilities. They are reported in Table 8.

Table 8. The estimated error probabilities for the final model for the data in Table 3. See page 425 for the symbols

Parameter	Estimate	s.e.
$\beta_1(D-)$	0.01	0.0037
$\beta_1(D+)$	0.05	0.0090
$\beta_2(D-)$	0.60	0.0247
$\beta_2(D+)$	0.37	0.0183
α_1	0.02	0.0029
$\alpha_2(C-,D-)$	0.51	0.0139
$\alpha_2(C+,D-)$	0.43	0.0442
$\alpha_2(C-,D+)$	0.39	0.0222
$\alpha_2(C+,D+)$	0.33	0.0338

Note: The α_2 parameters are not estimated independently but using the following restrictions:

$$\alpha_2(C+,D-) = \alpha_2(C-,D-) \times \varkappa,$$

$$\alpha_2(C+,D+) = \alpha_2(C-,D+) \times \varkappa.$$

The parameters $\alpha_2(C-,D-)$, $\alpha_2(C-,D+)$ and $\varkappa$ are estimated independently. The estimate of $\varkappa$ is 0.8465 with s.e. 0.0580.

The probabilities of police misclassification have a striking feature. The police only rarely report belt when there was no belt (α_1), or injury when there was no injury (β_1). On the other hand, the probability that the police report no injury when, in fact, there was an injury (β_2) is very high. Similarly the probability that the police report no belt when the belt was used (α_2) is high.

A second feature of the police misclassification is that the error probabilities concerning injury depend on the extent of the damage, but not on the sex of the driver. It is understandable that the police work with more accuracy concerning injury when the damage is high. Injuries are, perhaps, more obvious then. The probability that the police do not notice belt usage depends on both damage and sex. It is not intuitively clear why the underreporting of belt usage is less severe for female drivers. The underreporting is for women only 85 % of what it is for men in the corresponding damage category. The sex and the damage effects on the errors are additive on a logarithmic scale.

The following technical details about our model are relevant. Three cells had to be excluded from the fit. Reading across rows of Table 3, these are numbers 40, 55, and 56. The fitted frequencies are given in parentheses in Table 3. The fitted frequency for cell number 56 is only 0.05. Both the likelihood ratio and the Pearson test statistics react excessively to so small a denominator. The cell numbers 40 and 55 are more mysterious. They both belong to the 1974 data. Including them in the estimation and testing changes the overall fit to a poor one. The p-value drops from 0.24 to 0.09. They do not, however, change the parameter estimates much, with the notable exception of the estimate for the seat belt effect (A), which moves from −0.35 (0.080) to −0.49 (0.063). The basic structure of our model is thus resistant to the influence of these cells. We believe it is more reasonable to base inference on a well fitting model for 77 cells, than on almost the same model with a poor overall fit. We suspect data collection faults for cell numbers 40 and 55. Note, from Table 3, that the other fourteen 1974 cells also have large observed frequencies, and remarkably close fits.

The degrees of freedom are calculated as $80 - 8 - 4 - 13 - 3 = 52$. There are 80 cells in Table 3, the numbers of fixed (A, C, D) and (C, D) marginal totals are 8 and 4. The number of unrestricted parameters in Tables 7 and 8 is 13, and finally the three cells mentioned above were excluded.

The structural part of the model is relatively unaffected by the choice of error model. There was virtually no trade-off between modeling the errors more parsimoniously and including more interaction terms in the structural model. In general the final model has low correlations between the estimates. In particular the correlations between the structural estimates and the error estimates are low. The largest of the absolute values is 0.55 and the median is 0.09.

We estimated many error models. Espeland and Odoroff (1985, p. 667) suggest that there might be a dependence between police reporting of injury and seat belt usage. To study this we elaborated Table 6, letting the α_2 error probabilities depend on B^*, and letting the β_2 error probabilities depend on A^*. All models of this type were inferior to the one we report. Actually, most models of this type were poorly identified, with very flat likelihood functions close to the maximum.

Having established that the model fits well, we return to the interpretation of the structural part. From the estimates reported in Table 7 we calculate the corresponding probabilities of true injury given seat belt usage (A), sex (C), and damage (D). These are reported in Table 9, which should be compared to Table 4. It is fair to characterize the figures in Table 9 as the estimated probabilities of injury based on all

Table 9. The estimates of the probability of injury based on the data in Table 3, and adjusted for misclassifications in accordance with the fitted model. Cf. Table 4

		A–	*A*+
C–	*D*–	0.15	0.11
	D+	0.40	0.32
C+	*D*–	0.26	0.20
	D+	0.52	0.43

the data in Table 3, and adjusted for the contamination by misclassification of factors *A* and *B*.

The similarity rather than the dissimilarity between the figures in Tables 4 and 9 is striking. The belt effect (A) is significant in Table 9 while it was insignificant in Table 6. The same is true for the interaction ($C.D$) between sex and damage. On balance we dare say that the benefit of having access to doubly sampled data and a model for analyzing it is in this case twofold.

i. The singly classified data adds enough information to establish a significant seat belt effect. The power of the doubly classified data alone is insufficient for this.

ii. The doubly classified data permits us to model the probabilities of misclassification. Some of these are high enough to deserve serious attention.

5. Discussion

We want to call attention to a general feature of the example analyzed in Sections 3 and 4. The police report procedure is obviously set up primarily for administrative purposes. The question of whether the seat belt has an effect on the probability of injury is best characterized as a structural inference problem. Statistics collected for administrative purposes or as an offspring to administrative measures can be of great potential power for drawing structural inference. One should, however, adjust for errors of classification of measurement, or in general, nonsampling errors. The most natural way of finding information on the nonsampling errors is to set up a double sampling scheme.

Highway safety research is certainly not the only field of study in which large administrative registers are available and can be combined with information from small intensive surveys to improve accuracy. Public health, mentioned in the introductory example, criminology, and sociology are other fields for potential fruitful use of double sampling schemes. In fact, the idea of correcting official registers by personal interviews is not new. In Scandinavia population registers were established long before census practices. For decades, information from censuses has been used to correct the population registers.

With a more widespread use of double sampling the allocation problem needs attention. For what proportion of the sample should precise information be sought, when resources are limited and precise information is expensive. Here the costs of different types of information can be viewed broadly, including time and inconvenience aspects. Tenenbein (1970, 1972) discussed allocation rules for estimating a single binomial probability, and the probabilities of a multinomial distribution. Palmgren (1987) treated double sampling allocation for estimating the differences of proportions on the logit scale in two populations. The solution to the allocation problem is not known for more complex problems. In general, the optimal allocation depends on the specific form of the model under study, and on the true unknown values of the structural parameters and error probabilities. In any particular situation decisions could be based on simulated data made to mimic the structure to be investigated. It is worth noting that for some combinations of error probabilities and relative costs the solution is to devote all your resources to obtain precise measurements.

6. References

Dobson, A.J. (1983): An Introduction to Statistical Modelling. Chapman and Hall, London.

Chen, T.T., Hochberg, Y., and Tenenbein, A. (1984): Analysis of Multivariate Categorical Data with Misclassification Errors by Triple Sampling Schemes. Journal of Statistical Planning and Inference, 9, pp. 177–184.

Ekholm, A., Green, M., and Palmgren, J. (1986): Fitting Exponential Family Nonlinear Models in GLIM 3.77. Glim Newsletter, Issue 13, pp. 4–13.

Espeland, M.A. and Odoroff, C.L. (1985): Log-Linear Models for Doubly Sampled Categorical Data Fitted by the EM Algorithm. Journal of the American Statistical Association, 80, pp. 663–670.

Hochberg, Y. (1977): On the Use of Double Sampling Schemes in Analyzing Categorical Data with Misclassification Errors. Journal of the American Statistical Association, 72, pp. 914–921.

Nelder, J.A. and Wedderburn, R.W. (1972): Generalized Linear Models. Journal of the Royal Statistical Society, Series A, 135, pp. 370–384.

McCullagh, P. and Nelder, J.A. (1983): Generalized Linear Models. Chapman and Hall, London.

Palmgren, J. (1981): The Fisher Information Matrix for Log-Linear Models Arguing Conditionally on Observed Explanatory Variables. Biometrika, 68, pp. 563–566.

Palmgren, J. (1987): Precision of Double Sampling Estimators for Comparing Two Probabilities. Biometrika, 74, pp. 687–694.

Palmgren, J. and Ekholm, A. (1987): Exponential Family Nonlinear Models for Categorical Data with Errors of Observation. Applied Stochastic Models and Data Analysis, 3, pp. 111–124.

Payne, C.D., (Ed.) (1985): The GLIM System Release 3.77 Manual. NAG, Oxford.

Tenenbein, A. (1970): A Double Sampling Scheme for Estimating from Binomial Data with Misclassification. Journal of the American Statistical Association, 65, pp. 1350–1361.

Tenenbein, A. (1972): A Double Sampling Scheme for Estimating from Misclassified Multinomial Data with Applications to Sampling Inspection. Technometrics, 14, pp. 187–202.

Received April 1987
Revised October 1987