NEW FINDINGS ON THE USAGE OF LOGISTIC REGRESSION IN ACCIDENT DATA ANALYSIS

Jens-Peter Kreiss
Technische Universität Braunschweig
Germany
Tobias Zangmeister
Fraunhofer ITWM, Kaiserslautern
Germany
Paper Number 11-0192

ABSTRACT

In this paper we deal with different ways of statistical modeling of real world accident data in order to quantify the effectiveness of a safety function or a safety configuration (i.e. a specific combination of safety functions) in vehicles. It is shown that the effectiveness can be estimated along the so-called relative risk, even if the effectiveness does depend on a confounding variable, which may be categorical or continuous. In a second step the quite usual and from a statistical point of view classical logistic regression modeling is investigated. Main emphasis is laid on the understanding of the model and the interpretation of the occurring parameters. It is shown that the effectiveness of the safety function also can be detected via such a logistic approach and that relevant confounding variables can and should be taken into account. The interpretation of the parameters related to the confounder and the quantification of the influence of the confounder is shown to be rather problematic. All theoretical results are illuminated by numerical data examples.

INTRODUCTION

It is a relevant topic in accident research to quantify the possible effectiveness of a safety function or a safety configuration in passenger vehicles on the accident behavior. When dealing with a primary safety function, it is most relevant to determine the ability of this function to avoid accidents. In classical statistical theory one would assume that two different groups of vehicles can be observed over a certain period (e.g. one year) driving on the roads (experimental group and control group). The two groups are supposed to only differ according to whether the respective vehicles are equipped or not equipped with the safety function or safety configuration. Having observed the accident behavior, one could compare the two relative frequencies of having a specific type of accident in the two groups. To be a little bit more specific, we compare along the just described lines the two probabilities of having a (specific) accident given that the safety function is active or not. If we assume that for the random variable Z the event {Z=1} indicates that the accident of interest occurs, where S indicates whether the safety configuration is active (S=1) or not (S=0) and X denotes a further random variable (confounder) which may have some influence on the accident behavior and/or the safety equipment, we compare the following conditional probabilities.

$$P\left(Z=1 \mid S=r, X=x\right), r \in \{0,1\}, x \in \mathbf{X} \qquad (1)$$

Here **X** denotes the set of all possible outcomes of *X*. In applications *X* may be the gender of the driver, the age of the driver or of the vehicle, the mass of the vehicle or a selection (or even all) of these values as an example. So much for the pure statistical theory, in the real world one cannot carry out such an investigation by obvious reasons. The possible effectiveness of a safety function has to be quantified on the basis of accident data, only. This immediately implies that one cannot estimate the probability given in (1). If we extend the definition of the accident indicator *Z* as follows

$$Z = \begin{cases} 0, & \text{accident neutral to the safety function of interest} \\ 1, & \text{accident sensitive to the safety function of interest} \\ 2, & \text{no accident or accident not reported to database} \end{cases} \qquad (2)$$

then it is reasonable to assume that we can estimate the conditional probability

$$P\left(Z=1 \mid S=r, X=x, Z \in \{0,1\}\right), r \in \{0,1\}, x \in \mathbf{X} \quad (3)$$

only. The expression in (3) is a *conditional* probability which is indicated by "|" and quantifies the probability of the event *Z*=1 given that *S*=*r* (safety function active (*r*=1) or not (*r*=0)), given that we are in the subgroup described by the confounder *X*=*x* and given that an accident has occurred which has been reported to the underlying accident database and that this accident is *neutral* or *sensitive* to the safety function or safety configuration of interest ($Z \in \{0,1\}$).

However, in order to quantify a possible effectiveness of the safety function, we still are interested in the following ratio for $x \in \mathbf{X}$

$$RR(x) := \frac{P(Z=1 \mid S=1, X=x)}{P(Z=1 \mid S=0, X=x)}\ ,\ x \in \mathbf{X} \qquad (4)$$

which quantifies the performance of the safety function and is called *relative risk* in the following. The quantity

$$1 - RR(x) =: Eff(x)\ ,\ x \in \mathbf{X} \qquad (5)$$

is a measure of the effectiveness of the safety function for the group X=x, and describes the rate of accidents of interest within the group X=x which can be avoided by the safety function. It is shown in this paper that the relative risk as well as the effectiveness of a safety function or safety configuration reasonably can be estimated on the basis of accident data only. There is no conceptual difference between the cases where the confounder X is categorically or continuously distributed, as will be shown.

Of course many papers in the literature use a similar approach for quantifying the effectiveness of a safety function (cf. Tingvall et al. (2003), Martin et al. (2003), Dang (2004), Farmer (2004), Otto (2004), Page and Cuny (2004), Grömping et al. (2005) and Kreiss et al. (2005)). For a methodological overview concerning statistical methods applied to real-world accident data we refer to Hautzinger (2003), Grömping et al. (2007) and Hautzinger et al. (2008), while a complete statistical description of the logistic regression method can be found in Agresti (1996).

Many of the approaches rely on a logistic regression modeling of accident data, which not really is necessary for estimating $RR(x)$, cf. (4). The present paper discusses estimates for the relative risk $RR(x)$ and sheds some light on the interpretation of the parameters of a logistic regression when applied to accident counts. In principle there are at least two possibilities to introduce a logistic modeling to the situation of interest. From a classical statistical point of view one would be tempted to model the conditional probability of suffering an accident of interest, i.e.

$$P(Z=1 \mid S=r, X=x) = \frac{\exp(\alpha_0 + \alpha_1 \cdot r + \alpha_2 \cdot x)}{1+\exp(\alpha_0 + \alpha_1 \cdot r + \alpha_2 \cdot x)}, r \in \{0,1\}, x \in \mathbf{X} \qquad (6)$$

Here we assume for the sake of simple notation that X is univariate. Since we do not observe absolute numbers of traffic participants and following the discussion from above it may be more appropriate to use the logistic modeling in a different context as follows

$$P(Z=1 \mid S=r, X=x, Z \in \{0,1\}) = \frac{\exp(\beta_0 + \beta_1 \cdot r + \beta_2 \cdot x)}{1+\exp(\beta_0 + \beta_1 \cdot r + \beta_2 \cdot x)} \qquad (7)$$

i.e. modeling the conditional probability that an accident of interest occurs given that the safety function is on or off (S=1 or 0), that the confounder X takes the value x (e.g. a specific age of the vehicle) and given that an accident, which is *neutral* or *sensitive* to the safety function or safety configuration of interest has happened. Using the model (7) the typically wanted assertion

$$P(Z=1 \mid S=r, X=x, Z \in \{0,1\}) = \frac{1}{1+\exp(\beta_0 + \beta_1 \cdot r + \beta_2 \cdot x)} \qquad (8)$$

holds, which definitely is not the case for the modeling in (6) because the event Z=1 is not the complement of the event Z=0. To see this recall that the complement to the event that an accident of interest (i.e. sensitive to the safety function) has happened ($\{Z=1\}$) means a neutral accident ($\{Z=0\}$) *or* another accident *or* (and this by far is largest group) no accident (or a not reported accident) at all has happened ($\{Z=2\}$).

As it is argued above we need to get some information on the conditional probability $P\{Z=1 \mid S=r, X=x\}$ or more realistic about the ratio

$$\frac{P(Z=1 \mid S=1, X=x)}{P(Z=1 \mid S=0, X=x)} \qquad (9)$$

Later we will see what the implications of model (8) for (9) concerning this question are. Moreover it is of great interest what the interpretations of the parameters β_1 and β_2 (cf. model (7)) as well as α_1 and α_2 (cf. model (6)) are and how they relate to each other. So the main focus of the paper is to shed some light on the correct interpretation of results of (standard) logistic regression in accident analysis.

The paper is organized as follows. We start in the next section with an example from real-world accident data and continue in a further section with simulated accident data in order to be able to observe what the two different modelings ((6) and (7)) really measure. In simulated data we have the advantage that we really and exactly know what the underlying situation is. We continue in describing in detail the already mentioned

two different logistic regression modelings as well as their assumptions, consequences and interpretations. Finally we come back to our simulated accident data from the next but one section and apply the developed methodology to this data. There we will see whether and if yes to what extend we can estimate parameters of the two models.

REAL – WORLD ACCIDENT DATA EXAMPLE

Consider the following results obtained from real-world accident data collected within the German In Depth Accident Study (GIDAS). We focus on the quantification of the effectiveness of the electronic stability control (ESC) for passenger vehicles in Germany. From 12,833 recorded passenger vehicles involved in accidents, for which we know about the ESC-equipment and about the gender of the driver, a logistic regression can be carried through for the dependent variable

$$Z = \begin{cases} 0, & \text{accident neutral to ESC} \\ 1, & \text{skidding accident} \end{cases} \tag{10}$$

We have chosen the accident category *parking accident* as neutral to ESC, as we assume that ESC has no influence on the risk of suffering a parking accident. The observed data are condensed in the 2 × 2 contingency tables displayed in the Tables 1 and 2, separately for female and male drivers.

	ESC equipped	
Accident type	No	Yes
Parking accident	90	9
Skidding accident	387	9

Table 1: Real-world accident data for passenger cars with female driver

	ESC equipped	
Accident type	No	Yes
Parking accident	191	31
Skidding accident	782	38

Table 2: Real-world accident data for passenger cars with male driver

From Tables 1 and 2 one easily can compare the rates of ESC-equipment for the group of ESC-sensitive *skidding* accidents with the ESC-rates for the neutral accidents for the two gender categories. Doing so we obtain for male drivers a computed (crude) effectiveness of ESC of

$$Eff_{crude,male} = 1 - OR_{crude,male} = 1 - \frac{38 \cdot 191}{782 \cdot 31} = 0.701 = 70.1\% \tag{11}$$

and for female drivers of $Eff_{\text{crude, female}}$= 76.7%. We refer to the value $OR_{\text{male, crude}}$ = (38·191)/(782·31) = 0.299 as the crude Odds ratio for accidental situations with male drives and accordingly for female drivers ($OR_{\text{female, crude}}$ = 0.233). Adding all accidents in the four categories for male and female drivers we obtain a (crude) overall effectiveness of ESC of Eff_{crude}= 71.8%. For the calculation of standard odds-ratios we refer to Evans (1998) or Agresti (1996).

At this place we even do not want to stick to the absolute values of the effectiveness of ESC but to the fact that we obtain a 9.5% higher effectiveness of ESC in accidental situations in which the vehicle was driven by a woman. Rather we interpret the obtained result as an indication that we should include *gender of driver* as an explaining variable (confounder) into the logistic regression analysis. We expect of course a positive efficiency for both ESC-equipment and female drivers (compared to male drivers). Interestingly the results are not as expected. Standard software leads to the astonishing result that the coefficient for the variable *ESC (1=ESC on board)* is -1.260 (leading to an effectiveness of 71.6% but that the coefficient for the variable *Gender of Driver (1=female driver)* mounts to +0.032, leading to a *negative* effectiveness of -3.3% for female drivers. This is in contrast to the above results obtained when the accidents are considered separately for male and female drivers.

We refer to Kreiss et al. (2005), where a rather similar result of higher effectiveness of ESC for vehicles with female drivers has been described. There it is argued that the higher effectiveness of ESC in accidents with female drivers most likely is a pseudo-effect, which can be explained by a high correlation of gender of driver and size of vehicle. But this question is not a major point within this example and also within this paper.

In order to get an impression what is going on and what might go wrong we continue in the next section with simulated accident data from a quite simple model, which we will discuss later in detail in the section on logistic regression modeling Type II. It is necessary to consider simulated accident data because only in such a

case we are able to see what may happen and to thoroughly decide whether a suggested procedure works well or not.

SIMULATED DATA EXAMPLE

Let us assume that we have n=1,000,000 vehicles on the road. Further assume that 30 % of the vehicles are equipped with ESC. We think of *gender of driver* as a confounder X (X=1 refers to female and X=0 to male) and observe skidding accidents (i.e. Z=1) as accidents sensitive to ESC (accidents of interest) and some kind of neutral accidents (e.g. parking accidents) which refer to Z=0. Assume that the probability of suffering a loss of control accident for a passenger car is modeled according to the following logistic-type probability

$$P(Z=1 \mid S=r, X=x) = \frac{\exp(\beta_0 + \beta_1 \cdot r + \beta_2 \cdot x)}{1+\exp(\beta_0 + \beta_1 \cdot r + \beta_2 \cdot x)} \qquad (12)$$

for all $r,x \in \{0,1\}$ and as parameters we choose

$$\beta_0 = -5.0 \quad \beta_1 = -0.35 \quad \beta_2 = +0.50 \qquad (13)$$

This means that we assume a rather high positive effectiveness of ESC as well as a positive effectiveness of gender equal to male on the risk of suffering a skidding accident. From the above settings we obtain Table 3, showing the probabilities for suffering a skidding accident when driving a certain period, e.g. one year, on the roads. Of course these probabilities have to be rather small, since accidents are rare events.

	Gender	
ESC equipped	male ("0")	female ("1")
No ("0")	$6.93\ 10^{-3}$	$1.11\ 10^{-2}$
Yes ("1")	$4.73\ 10^{-3}$	$7.77\ 10^{-3}$

Table 3: Probabilities for a skidding accident

The assumption (12) not really coincides with the typical binary logistic regression modeling for accident data. There typically the conditional probability

$$P(Z=1 \mid S=r, X=x, \text{a reported accident has happened}) \qquad (14)$$

is modeled by the expression given on the right hand side of (12). This really makes a difference and we will discuss this point later in detail.

We further assume that 80% of the vehicles are driven by male drivers. The exact distribution of male and female drivers within ESC-equipped and non-equipped vehicles is as follows.

	Gender		
ESC equipped	0	1	sum
0	600,000	100,000	700,000
1	200,000	100,000	300,000
sum	800,000	200,000	1,000,000

Table 4: Driver distribution in ESC-equipped and non-equipped vehicles

Table 4 reflects that 30% of the vehicles are equipped with ESC and shows that 50% of the females drive an ESC-equipped vehicle and only 25% of the males drive an ESC-equipped vehicle. All these values refer to exposure data (vehicles on the road) and not accidents.

According to our assumption we obtain by Monte Carlo simulation from the probabilities of Table 3 the accident counts displayed in Table 5.

	Gender		
ESC equipped	0	1	sum
0	4,009	1,097	5,106
1	951	779	1,730
sum	4,960	1,876	6,836

Table 5: Simulated numbers of skidding accidents "Z=1"

Concerning the neutral accidents we consider the two scenarios shown in Tables 6 and 7.

Scenario I (cf. Table 6) rather accurately resembles the underlying exposure distribution (cf. Table 4) according to equipment with ESC and gender of the driver. Scenario II (cf. Table 7) accurately resembles the ESC-

equipment distribution within the two gender groups (compare the distribution within the columns of Tables 4 and 7) but the probability of suffering a neutral accident varies with the gender of the driver.

	Gender		
ESC equipped	0	1	sum
0	5,760	960	6,720
1	1,920	960	2,880
sum	7,680	1,920	9,600

Table 6: Neutral accidents "Z=0" (scenario I)

and

	Gender		
ESC equipped	0	1	sum
0	4,050	2,100	6,150
1	1,350	2,100	3,450
sum	5,400	4,200	9,600

Table 7: Neutral accidents "Z=0" (scenario II)

Using the SPSS-routine *logistic regression* the following estimates for scenario I (i.e. skidding accidents according to Table 5 and neutral accidents according to Table 6) are obtained:

$$\hat{\beta}_0^I = -0.362 \quad \hat{\beta}_1^I = -0.341 \quad \hat{\beta}_2^I = +0.495 \qquad (15)$$

The estimated coefficient $\hat{\beta}_1^I$ and $\hat{\beta}_2^I$ perfectly match the underlying situation, cf. (13). However the estimator $\hat{\beta}_0^I$ is not consistent. This is not surprising because this value mainly controls the absolute value of the corresponding probability in (12) and this is not comparable with the relative frequencies within the group of accidents only.

The results for scenario II, i.e. skidding accidents according to Table 5 and neutral accidents according to Table 7, read as follows

$$\hat{\beta}_0^{II} = -0.010 \quad \hat{\beta}_1^{II} = -0.341 \quad \hat{\beta}_2^{II} = -0.640 \qquad (16)$$

It can be seen that $\hat{\beta}_1^{II}$ still works rather well, but $\hat{\beta}_2^{II}$ does not. Why this is the case will be discussed in a later section of this paper.

Finally let us see what happens within our two data scenarios I and II when we apply the logistic regression routine without taking the gender of the driver as a confounding variable into account. Then we come up with simple 2 × 2 contingency tables (cf. Tables 8 and 9)

	Accident	
ESC equipped	neutral ("0")	skidding ("1")
0	6,720	5,106
1	2,880	1,730

Table 8: Simulated accident data according to scenario I

	Accident	
ESC equipped	neutral ("0")	skidding ("1")
0	6,150	5,106
1	3,450	1,730

Table 8: Simulated accident data according to scenario II

The estimators for the effectiveness of ESC in the merged situation and without any confounding variable are rather easily computed, cf. Evans (1998) or Agresti (1996), and read as follows

$$1 - Eff_I = 0.791 \quad \text{and} \quad 1 - Eff_{II} = 0.604 \qquad (17)$$

It can be seen that both values substantially differ from the underlying effectiveness of

$$1 - Eff_{Model} = 0.705 \qquad (18)$$

This demonstrates that it is essential to include a confounding variable when there is one with a non-negligible influence.

LOGISTIC MODELING TYPE I

In this section we deal with the following logistic regression modeling for the probability of suffering an accident of interest given the states of the safety function, the value of the confounder and the fact that an accident of interest or a neutral accident has happened. To be precise we assume

$$P\left(Z=1 \mid S=r, X=x, Z \in \{0,1\}\right) = \frac{\exp\left(\beta_0+\beta_1 \cdot r+\beta_2 \cdot x\right)}{1+\exp\left(\beta_0+\beta_1 \cdot r+\beta_2 \cdot x\right)} \tag{19}$$

for $r \in \{0,1\}$, $x \in \mathbf{X}$. β_0, β_1 and β_2 denote the parameters of the model.

We emphasize that the conditional probability in (19) varies not only in r and x (the status of the safety function and the confounder) but also with the random event $Z \in \{0,1\}$. This means for example that if the probability of suffering an accident of neutral type changes, then the modeled conditional probabilities will vary as well. This already explains that the interpretation of the coefficients β_1 and β_2 really is delicate.

(19) is equivalent to assume

$$\ln\left(\frac{P\left(Z=1 \mid S=r, X=x, Z \in \{0,1\}\right)}{1-P\left(Z=1 \mid S=r, X=x, Z \in \{0,1\}\right)}\right) = \beta_0+\beta_1 \cdot r+\beta_2 \cdot x \tag{20}$$

i.e. a linear relationship of the logit (the left hand side of (20)) on the values r and x of S and X, respectively. For later reference we state here that (21) holds true.

$$1-P\left(Z=1 \mid S=r, X=x, Z \in \{0,1\}\right) = P\left(Z=0 \mid S=r, X=x\right) \ \forall x \in \mathbf{X}, r \in \{0,1\} \tag{21}$$

Standard statistical software now easily allows to compute estimators $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ from observations (Z_k,S_k,X_k), k=1, ... , n. Such observations typically are provided from accident databases.

The main question now is, how one can interpret the parameters β_0, β_1 and β_2. To receive some results in this direction observe

$$\begin{aligned} &P\left(Z=1 \mid S=r, X=x\right) \\ &= P\left(Z=1 \mid S=r, X=x, Z \in \{0,1\}\right) \\ &\cdot \frac{P\left(S=r, X=x, Z \in \{0,1\}\right)}{P\left(S=r, X=x\right)} \end{aligned} \tag{22}$$

Since

$$\begin{aligned} &\frac{P\left(S=r, X=x, Z \in \{0,1\}\right)}{P\left(S=r, X=x\right)} \\ &= \frac{P\left(S=r, X=x, Z=1\right)}{P\left(S=r, X=x\right)} + \frac{P\left(S=r, X=x, Z=0\right)}{P\left(S=r, X=x\right)} \\ &= P\left(Z=1 \mid S=r, X=x\right) + P\left(Z=0 \mid S=r, X=x\right) \end{aligned} \tag{23}$$

one obtains from (23) and (19)

$$\begin{aligned} &P\left(Z=1 \mid S=r, X=x\right) = P\left(Z=1 \mid S=r, X=x, Z \in \{0,1\}\right) \\ &\cdot\left(P\left(Z=1 \mid S=r, X=x\right)+P\left(Z=0 \mid S=r, X=x\right)\right) \\ &\Leftrightarrow \\ &P\left(Z=1 \mid S=r, X=x\right) \\ &\cdot\left(1-P\left(Z=1 \mid S=r, X=x, Z \in \{0,1\}\right)\right) \\ &= P\left(Z=1 \mid S=r, X=x, Z \in \{0,1\}\right) \\ &\cdot P\left(Z=0 \mid S=r, X=x\right) \end{aligned} \tag{24}$$

and therefore

$$\begin{aligned} &P\left(Z=1 \mid S=r, X=x\right) \\ &= \exp\left(\beta_0+\beta_1 \cdot r+\beta_2 \cdot x\right) \cdot P\left(Z=0 \mid S=r, X=x\right) \end{aligned} \tag{25}$$

(25) looks rather similar to a logistic regression model for the conditional probability $P(Z=1|S=r,X=x)$, but it is not! To see this observe that $P(Z=0|S=r,X=x) \neq 1-P(Z=1|S=r,X=x)$ because Z also can take the value 2, which stands for the event "'no accident or accident not reported to data base'". The just stated inequality does not even hold approximately, since both probabilities - in contrast to $P(Z=2|S=r,X=x)$ - typically are extremely small. But the following essential equality is true

$$\frac{P\left(Z=1 \mid S=1, X=x\right)}{P\left(Z=1 \mid S=0, X=x\right)} = e^{\beta_1} \cdot \frac{P\left(Z=0 \mid S=1, X=x\right)}{P\left(Z=0 \mid S=0, X=x\right)} \tag{26}$$

for all $x \in \mathbf{X}$.

For further calculations we need the following essential assumption

Assumption A1: Assume that the events "S=r", r ∈ {0,1}, and "Z=0" are independent (given that X=x holds).

(A1) implies that P(S=r|Z=0,X=x)=P(S=r|X=x) for x∈ {0,1} and x ∈ **X**. (26) leads under assumption (A1) and because of

$$\begin{aligned} &P(Z=0\mid S=r,X=x)\\ &=\frac{P(Z=0,S=r,X=x)}{P(X=x)}\cdot\frac{P(X=x)}{P(S=r,X=x)}\\ &=\frac{P(Z=0,S=r\mid X=x)}{P(S=r\mid X=x)}\\ &=P(Z=0\mid X=x),\ \text{by A1 for } r=1,2 \end{aligned} \tag{27}$$

immediately to

$$1-Eff(x)=\frac{P(Z=1\mid S=1,X=x)}{P(Z=1\mid S=0,X=x)}=e^{\beta_1} \tag{28}$$

Thus it has been shown that a logistic regression modeling (19) on the accident level leads under the reasonable assumption (A1) to a constant relative risk or effectiveness of the safety function in dependence of the confounder value *x*. The logistic regression approach (19) does not allow for a relative risk or effectiveness of a safety function which varies with the value *x* of the confounding variable *X*. For a method which allows for relative risk and effectiveness of the safety function which may vary with the value x of the confounding variable X we refer to Kreiss and Zangmeister (2011).

A remaining question still is how one shall interpret β_0 and β_2. Since there is no hope of interpreting β_0, the question is whether β_2 describes the influence of the confounding variable *X* not only for the conditional probability *P*(*Z*=1 | *S*=*r*,*X*=*x*,*Z* ∈ {0,1}) on the accident level but also for the conditional probability *P*(*Z*=1 | *S*=*r*,*X*=*x*) of interest. One might be tempted to assume that this indeed is true. We will investigate this question in the following. To do so we assume within the model (19) that the confounding variable *X* is categorical and takes the values 0 and 1, only.

From the key equation (25) one obtains for *r* ∈ {0,1}

$$\frac{P(Z=1\mid S=r,X=1)}{P(Z=1\mid S=r,X=0)}=e^{\beta_2}\cdot\frac{P(Z=0\mid S=r,X=1)}{P(Z=0\mid S=r,X=0)} \tag{29}$$

Now for *r*, *x* ∈ {0,1} and if assuming (A1)

$$\begin{aligned} &P(Z=0\mid S=r,X=x)\\ &=\frac{P(Z=0,S=r,X=x)}{P(X=x)}\cdot\frac{P(X=x)}{P(S=r,X=x)}\\ &=\frac{P(Z=0,S=r\mid X=x)}{P(S=r\mid X=x)}\\ &=P(Z=0\mid X=x) \end{aligned} \tag{30}$$

Thus one obtains for *r* ∈ {0,1}

$$\frac{P(Z=1\mid S=r,X=1)}{P(Z=1\mid S=r,X=0)}=e^{\beta_2}\cdot\frac{P(Z=0\mid X=1)}{P(Z=0\mid X=0)} \tag{31}$$

which immediately leads to the following formula

$$\frac{\dfrac{P(Z=1|S=1,X=1)}{P(Z=1|S=1,X=0)}}{\dfrac{P(Z=1|S=0,X=1)}{P(Z=1|S=0,X=0)}}=1 \tag{32}$$

(32) means that the ratio of probabilities of having an accident of type of interest given *X*=1 or *X*=0, when driving on the roads, does not vary with having the safety function on board or not. Still the confounder very well may have some influence on the risk of suffering an accident of interest.

β_2 describes the difference of the relative risk of having a neutral accident and the relative risk of having an accident of interest with or without the safety function active for the two groups *X*=0 and *X*=1. For example β_2=0 means that there is no difference in the relative risks for neutral or relevant accidents. Even so there still may be a significant influence of the confounding variable on the probabilities of suffering a neutral or a relevant accident themselves.

Let us state another assumption:

Assumption A2: Assume that the conditional probability of suffering an accident of interest for any specific given value *X*=*x* is independent of the value *x*, i.e.

$$P(Z=0\mid X=x)\ \text{is independent of}\ x\in\mathbf{X} \tag{33}$$

With this assumption one obtains from (31) that

$$\frac{P(Z=1 \mid S=r, X=1)}{P(Z=1 \mid S=r, X=0)} = e^{\beta_2} \qquad (34)$$

which may be regarded as the typically interpretation for β_2, cf. formula (28).

It is common in literature to interpret β_2 according to (34) as the influencing 'effect' of the confounding variable X without further thoughts on the plausibility of assumption (A2), like described in the introductory example.

The question is whether or not assumption (A2) is reasonable. At first it can be seen that assumption (A2) is equivalent to

Assumption A3: Assume that the events "Z=0" and "X=x" are for all $x \in \mathbf{X}$ independent which may be expressed with

$$P(Z=0, X=x) = P(Z=0) \cdot P(X=x) \qquad (35)$$

This means that the category of neutral accidents is not only neutral concerning the safety function but also neutral according to the confounding variable. In other words assumption (A2) or equivalently assumption (A3) assumes that the probability of suffering a neutral accident is the same for all subgroups X=x, $x \in \mathbf{X}$, described by the confounder. This seems to be hardly justifiable and therefore the above interpretation of β_2 is more than doubtful. Thus, one has to stay with (31) and interpret β_2 according to (31).

Hence, there is really a difference in interpreting the parameters β_1 (cf. (28)) and β_2, cf. (31).

LOGISTIC MODELING TYPE II

A different and also possible modeling is to deal with conditional probabilities like

$$P(Z=1 \mid S=r, X=x), \quad r \in \{0,1\}, x \in \mathbf{X} \qquad (36)$$

directly and not additionally to condition on the event $Z \in \{0,1\}$ that an accident of neutral or relevant type has occurred. E.g. to assume a logistic regression model of the following form

$$P(Z=1 \mid S=r, X=x) = \frac{\exp(\alpha_0 + \alpha_1 \cdot r + \alpha_2 \cdot x)}{1+\exp(\alpha_0 + \alpha_1 \cdot r + \alpha_2 \cdot x)} \qquad (37)$$

for $r \in \{0,1\}$ and $x \in \mathbf{X}$.

The conditional probability in (37) in contrast to the conditional probability (19) does not vary with the random event $Z \in \{0,1\}$ and therefore does not vary with changing probabilities of suffering an accident of neutral type. This indicates that the interpretation of the coefficients α_1 and α_2 might be easier compared to the coefficients β_1 and β_2 in model (19).

Of course in this situation (and this again is in contrast to model (19)) we do have

$$P(Z=0 \mid S=r, X=x) + P(Z=1 \mid S=r, X=x) \neq 1 \qquad (38)$$

This implies that

$$P(Z=0 \mid S=r, X=x) \neq \frac{1}{1+\exp(\alpha_0 + \alpha_1 \cdot r + \alpha_2 \cdot x)} \qquad (39)$$

Note that both probabilities in (37) and (39) typically are extremely small and not even approximately add up to one!

Assume for example that the probability of having an accident of relevant type, i.e. Z=1 within a certain period (e.g. one year), is about 10^{-3} or lower then we have

$$\begin{aligned} P(Z=1 \mid S=r, X=x) &= \frac{\exp(\alpha_0 + \alpha_1 \cdot r + \alpha_2 \cdot x)}{1+\exp(\alpha_0 + \alpha_1 \cdot r + \alpha_2 \cdot x)} \qquad (40) \\ &\approx \exp(\alpha_0 + \alpha_1 \cdot r + \alpha_2 \cdot x) \end{aligned}$$

where the approximation is the better the lower the probability on the left hand side of (40) is.

A big advantage of model (37) is the interpretability of the parameters α_1 and α_2. Using the approximation in (40) one easily obtains

$$\frac{P(Z=1 \mid S=1, X=x)}{P(Z=1 \mid S=0, X=x)} \approx \mathrm{e}^{\alpha_1} \qquad (41)$$

which of course is much in line with the result (28) which has been obtained from model (19) only under the additional assumption (A1). However this does not seem crucial since we need (A1) at least for an estimate of the right hand side of (41) on the basis of accident data. Moreover we similarly obtain for any $x_0, x_1 \in \mathbf{X}$

$$\frac{P(Z=1 \mid S=r, X=x_1)}{P(Z=1 \mid S=r, X=x_0)} \approx \mathrm{e}^{\alpha_2} \qquad (42)$$

But again, if we intend to estimate the left hand side of (42), which equals

$$\begin{aligned}
&\frac{P(Z=1,S=r,X=x_1)}{P(Z=1,S=r,X=x_0)}\cdot\frac{P(S=r,X=x_0)}{P(S=r,X=x_1)}\\
&=\frac{P(Z=1,S=r,X=x_1)}{P(Z=1,S=r,X=x_0)}\cdot\frac{P(S=r\mid X=x_0)}{P(S=r\mid X=x_1)}\\
&\cdot\frac{P(Z=0\mid X=x_0)}{P(Z=0\mid X=x_1)}\cdot\frac{P(Z=0\mid X=x_1)}{P(Z=0\mid X=x_0)}\cdot\frac{P(X=x_0)}{P(X=x_1)}\\
&=\frac{P(Z=1,S=r,X=x_1)}{P(Z=1,S=r,X=x_0)}\cdot\frac{P(Z=0,S=r,X=x_0)}{P(Z=0,S=r,X=x_1)}\\
&\cdot\frac{P(Z=0\mid X=x_1)}{P(Z=0\mid X=x_0)},\quad \text{by A1}
\end{aligned}\tag{43}$$

and therewith

$$\begin{aligned}
e^{\alpha_2}\approx&\frac{P(Z=1,S=r,X=x_1)}{P(Z=1,S=r,X=x_0)}\\
&\cdot\frac{P(Z=0,S=r,X=x_0)}{P(Z=0,S=r,X=x_1)}\cdot\frac{P(Z=0\mid X=x_1)}{P(Z=0\mid X=x_0)}
\end{aligned}\tag{44}$$

We need a kind of assumption (A2) in order to have that the last factor in the equation above is know (e.g. equal to one). Note that the first two ratios easily can be estimated from accident data. Since it has been argued that assumption (A2) hardly is justifiable, we run into exactly the same problem following both ways of modeling. Here within the estimation of α_2, the term $P(Z=0|X=x_1)/P(Z=0|X=x_0)$ occurs which causes problems and in the modeling following assumption (19) exactly the same term causes difficulties in the interpretation of the parameter β_2, cf. (31).

Summarizing one can say that there are no big differences between the two modelings (19) and (37). The difficulties demanding for some further assumptions are nearly the same. Only the estimation procedures within the preceding section seem to be more standard since it is a modeling of the actual data and therefore usual statistical software packages likes SPSS, SAS or R can be used to compute parameter estimates. This is the reason why the modeling and results of the preceding section are recommended to be used.

SIMULATED DATA EXAMPLE – DISCUSSION

In the simulated data example section we introduced an example with simulated data, where the a priori known effectiveness of ESC was tried to be computed with a logistic regression. Two different scenarios were considered.

Scenario I (cf. Table 6) rather accurately resembled the underlying exposure distribution (cf. Table 4) according to equipment with ESC and gender of the driver. Scenario II (cf. Table 7) accurately resembled the ESC-equipment distribution within the two gender groups (compare the distribution within the columns of Tables 4 and 7) but the probability of suffering a neutral accident varies with the gender of the driver. Summarizing one can say that the data according to scenario I fulfilled the requirements given in assumptions (A1) and (A2) and the data according to scenario II only fulfilled (A1) but not (A2). Both scenarios I and II do not fulfill (A3).The results of the logistic regression were:

$$\begin{aligned}
&\hat\beta_0^I=-0.362 \quad \hat\beta_1^I=-0.341 \quad \hat\beta_2^I=+0.495\\
&\hat\beta_0^{II}=-0.010 \quad \hat\beta_1^{II}=-0.341 \quad \hat\beta_2^{II}=-0.640
\end{aligned}\tag{45}$$

compared to the a priori given model parameters

$$\beta_0^I=-5 \quad \beta_1^I=-0.35 \quad \beta_2^I=+0.5 \tag{46}$$

The estimated coefficients $\hat\beta_1^I$ and $\hat\beta_2^I$ perfectly match the underlying situation, cf. (46). However the estimator $\hat\beta_0^I$ is not consistent, which was already discussed in the section containing the real-world accident data example.

The estimated coefficient $\hat\beta_1^{II}$ still works rather well, but $\hat\beta_2^{II}$ does not.

Here one has to recall that the sufficient condition for the reliability of the estimator $\hat\beta_2^{II}$ is that $P(Z=0|X=x)$ is independent of x (cf. assumption (A2)). This is obviously the case in scenario I but not in scenario II as can be seen when looking for the two scenarios at the ratio $P(Z=0|X=1)/P(Z=0|X=0)$:

$$\frac{P(Z^I=0\mid X=1)}{P(Z^I=0\mid X=0)}=1 \tag{47}$$

and

$$\frac{P(Z^{II}=0\mid X=1)}{P(Z^{II}=0\mid X=0)}\approx 3.1 \tag{48}$$

The two different scenarios demonstrate that the effectiveness of the safety function reliably can be

estimated from accident data but that one has to be cautious with the estimators of the coefficients of the confounding variables.

Summarizing one can say that the effectiveness of a safety function reliably can be estimated in praxis, but that the influence of a confounder can hardly be quantified in general. Nevertheless it is rather essential to include relevant confounders in the investigation in order to quantify the (pure) effectiveness of a safety function correctly.

CONCLUSIONS

We have studied two different approaches of logistic regression modeling for accident data. It has been shown that in both cases and especially for the much easier to interpret model (6) standard logistic regression software leads not to absolutely exact but to rather reasonable estimators for the effectiveness of a safety function or safety configuration in vehicles under mild assumptions. Thus it has been shown that the effectiveness of a safety function or configuration reliably can be estimated in praxis. Concerning the possible influence of one or more confounders it is obtained that the corresponding effects hardly can be quantified in general. This is only possible under assumptions, which typically are not met in praxis. But it is extremely essential to include relevant confounders in the logistic regression investigation in order to quantify the effectiveness of a safety function correctly. This means that the effects of the confounders on the accident outcomes (which itself typically cannot be quantified!) does not lead to a bias in the quantification of the pure effectiveness of the safety function or configuration.

Concerning the presented real world accident data this means that we cannot rely on the estimated effectiveness of the confounder *gender of driver* on the risk of suffering a skidding accident (recall that we obtained from the logistic regression with that confounder a surprising negative effectiveness for female drivers) but we can rely on the calculated effectiveness of 71.6% for the ESC in this situation.

ACKNOWLEDGEMENT

The authors would like to acknowledge the EU commission, which supported the TRACE project, which gave birth to the main ideas behind this paper (cf. Zangmeister et al. (2008)). Especially we would like to sincerely thank Yves Page for countless fruitful discussions.

REFERENCES

[1] Aga M., Okada, A. (2003). Analysis of Vehicle Stability Control (VSC)'s Effectiveness From Accident Data. ESV-paper 541. 18th-ESV-Conference, Nagoya (Japan).

[2] Agresti, A. (1996). An Introduction to Categorical Data Analysis. Wiley Series in probability and Statistics (New York).

[3] Bahouth. G. (2005). Real World Crash Evaluation of Vehicle Stability Control (VSC) Technology. 49th Annual Proceedings of the Association for the Advancement of Automotive Medicine.

[4] Dang, J. N. (2004). Preliminary Results Analyzing the Effectiveness of Electronic Stability Control (ESC) Systems. U.S. Department of Transportation – National Highway Traffic Safety Administration. Evaluation Note.

[5] Evans, L. (1998). Antilock brake systems and risk of different types of crashes in traffic. ESV-paper No. 98-S2-O-12. 16th ESV-Conference, Windsor (Canada).

[6] Farmer, C.M. (2004). Effect of Electronic Stability Control on Automobile Crash Risk. Traffic Injury Prevention **5**, 317 – 325.

[7] Grömping, U., Menzler, S. and Weimann, U. (2005). Split-Register Study: A New Method for Estimating the Impact of Rare Exposure on Population Accident Risk based on Accident Register Data. 1st International ESAR Conference. Hanover (Germany) and Berichte der Bundesanstalt für Straßenwesen, sub series Fahrzeugtechnik, Report F55, 95 – 101.

[8] Grömping, U., Pfeiffer, M. and Stock, W. (2007). Statistical Methods for Improving the Usability of Existing Accident Databases. Deliverable 7.1 for EU-Project TRaffic Accident Causation in Europe (Project No. 027763 - TRACE).

[9] Hautzinger H. (2003). Measuring the Effect of Primary Safety Devices on Accident Involvement Risk of Passenger Cars – Some Methodological Considerations. Working paper of the. SARAC II project, Heilbronn, IVT-paper.

[10] Hautzinger, H., Grömping, U., Kreiss, J.–P., Mougeot, M., Pastor, C., Pfeiffer, M. and

Zangmeister, T. (2008). Summary Report of Work Package 7 – Statistical Methods, Deliverable 7.5 for EU–Project TRaffic Accident Causation in Europe (Project No. 027763–TRACE).

[11] Kreiss, J.-P., Schüler, L. and Langwieder, K. (2005). The Effectiveness of Primary Safety Features in Passenger Cars in Germany. ESV-paper No. 05-145. 19th -ESV -Conference, Washington D.C. (USA).

[12] Martin J.-L., Derrien Y., Laumon B. (2003). Estimating Relative Driver Fatality and Injury Risk According to Some Characteristics of Cars and Drivers Using Matched-Pair Multivariate Analysis. ESV-paper No. 364. 18th ESV-Conference, Nagoya (Japan).

[13] Otto, S. (2004). Quantifizierung des Einflusses aktiver Sicherheitssysteme auf die Unfallwahrscheinlichkeit und Identifikation von sicherheitsrelevanten Attributen basierend auf Realunfalldaten. Diploma Thesis, Universität Dortmund.

[14] Page Y., Cuny, S. (2004). Is ESP effective on French Roads. 1st -International ESAR, Hanover (Germany).

[15] Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis}. Chapman and Hall, London.

[16] Tingvall, C., Krafft, M., Kullgren, A., Lie, A. (2003). The Effectiveness of ESP (Electronic Stability Program) in Reducing Real Life Accidents. ESV -paper 261. 18th ESV-Conference, Nagoya (Japan).

[17] Weekes, A. Avery, M., Frampton, R. and Thomas, P. (2009). ESC Standard Fitment and Failure to Protect Young Drivers. Paper. ESV-paper No. 09-278. 21st ESV -Conference, Stuttgart (Deutschland).

[18] Zangmeister, T., Kreiss, J. –P., Schüler, L. (2008). Evaluation of Existing Safety Features. Deliverable 7.4 for EU–Project TRaffic Accident Causation in Europe (Project No. 027763 – TRACE).

[19] Kreiss, J. –P., Zangmeister, T., (2011). Quantification of the Effectiveness of a Safety Function in Passenger Vehicles on the Basis of Real-World Accident Data. Technical Report Fraunhofer ITWM No. 203