The use of prediction models for eliminating effects due to regression-to-the-mean in road accident data

Ulf Brüde and Jörgen Larsson

Reprint from Accident Analysis & Prevention, Vol. 20, No. 4, pp. 299-310, 1988

THE USE OF PREDICTION MODELS FOR ELIMINATING EFFECTS DUE TO REGRESSION-TO- THE-MEAN IN ROAD ACCIDENT DATA

ULF BRÜDE and JÖRGEN LARSSON

Swedish Road and Traffic Research Institute (VTI), S-58101 Linköping, Sweden

(Received 20 March 1987; in revised form 2 November 1987)

Abstract—In recent years, various methods have been proposed for estimating the true accident level, i.e. the expected number of accidents m when a total of x accidents have been observed at a junction, road section, etc., during a certain period of time. One such method has been named the Empirical Bayes Method (EB method). A description is given of a variant of the EB method utilizing prediction models for the number of accidents. Input data to the prediction models may consist, for example, of traffic flows in a junction. According to empirical comparisons of accidents in junctions, this variant of the EB method may be preferable in certain cases to the conventional EB method. However, it has not yet been determined how this variant of the EB method should generally take into account the precision of the prediction models. This means, for example, that in a nonexperimental before-and-after study of the effect of a particular action, varying results may be obtained according to the assumptions made concerning the precision of the prediction model.

INTRODUCTION

In planning measures to improve traffic safety, it is normal and in itself reasonable to select objects (such as junctions) where there are high numbers of accidents. But it has also been found that the observed accident numbers do not reflect true accident levels in this context. For the objects most affected by accidents, the number of accidents, is on the average an overestimate of the true accident level. Conversely, the opposite applies in the case of the least affected objects.

If we restrict ourselves to the choice of objects with large numbers of accidents, the above often implies that cost-benefit calculations for conceivable measures are too favourable and that subsequent before-and-after studies provide overestimates of the accident reduction effect afforded by the measures. This assumes that no attempt is made to estimate the true number of accidents instead of using the observed number of accidents (during the before period).

The phenomenon by which a randomly large number of accidents for a certain object during a before period is normally followed by a reduced number of accidents during a similar after period, even if no measures have been implemented (while the opposite applies in the case of a randomly small number of accidents) is generally termed regression-to-the-mean.

In recent years various methods [Abbess, Jarrett and, Wright, 1981; Hauer, 1986; Brüde and Larsson, 1982, 1985; Danielsson, 1986; Junghard and Danielsson, 1986] have been proposed for estimating the true accident level, i.e. the expected number of accidents m , when x accidents have been observed during a certain period of time.

It is desirable to have a method that can be used for individual objects and which compensates for both randomly large and randomly small accident numbers.

One of the proposed methods is named the Empirical Bayes Method (EB method). This report describes a number of ideas and results concerning one variant of the EB method where prediction models for the number of accidents are used.

THE EB METHOD

Assume a population of junctions, for example, with an expected number of accidents which can be regarded as observations of a stochastic variable. According to the

EB method [Abbess *et al.*, 1981; Hauer, 1986] the following will apply in the case of an individual junction:

$$E(m|x) = x + \frac{E(X)}{V(X)} [E(X) - x] \quad (1)$$

and that

$$\hat{m} = x + \frac{\bar{x}}{s^2} (\bar{x} - x) \quad (2)$$

where $\hat{m}$ is an estimate of the true accident level, i.e. the expected number of accidents m when x accidents have been observed. $\bar{x}$ and s^2 are the mean and variance of the number of accidents in the population from which the junction has been chosen.

The EB method implies that corrections of the observed values consequently are made in the direction of $\bar{x}$ for each unit observed.

In practical applications, several significant problems may occur: (i) It may be difficult to define the population in a suitable way. Ideally, the population should consist only of junctions of the same type as the junction for which the expected number of accidents m is to be estimated; (ii) It may be difficult to obtain accident data for the defined population; (iii) If the population is small, $\bar{x}$ and s^2 will be uncertain estimates.

A PERFECT PREDICTION MODEL

Assume a perfect prediction model for the number of accidents. In (2), $\bar{x}$ would then be replaced by the predicted value (= "*pred*") and also s^2 by "*pred*" (assuming a Poisson distribution) which would mean that:

$$\hat{m} = x + \frac{pred}{pred} (pred - x) = pred. \quad (3)$$

In practice, it is impossible to achieve perfect prediction models. In the case of junctions, models of the following type are commonly used:

$$pred = f(\text{traffic flows}).$$

If junctions with larger numbers of accidents than predicted are chosen, it will be true on the average that:

$$pred < m < x.$$

If instead, junctions with smaller numbers of accidents than predicted are chosen, it will be true on the average that:

$$x < m < pred.$$

THE EB METHOD USING NONPERFECT PREDICTION MODELS

Assume a number of objects of the "same" type (e.g. junctions with the same number of legs and similar traffic flows). For each of these junctions, the predicted number of accidents is equal to "*pred*." Although the junctions are of the "same" type, there are nevertheless differences that cannot be taken into account in the prediction models. This means that the true expected number of accidents m varies somewhat from junction to junction.

Assume now that m can be regarded as a gamma-distributed stochastic variable* with an expected value equal to “ $pred.$ ” The following will then apply:

$$E(m) = \frac{q}{p} = pred \quad (4)$$

and

$$V(m) = \frac{q}{p^2} = \frac{pred}{p}. \quad (5)$$

If the prediction model were perfect, $V(m) = \frac{pred}{p} = 0$ and the entire probability mass for the gamma distribution would be concentrated to “ $pred.$ ”

For a given junction among junctions of the “same” type, it may be assumed that the number of accidents X has a Poisson distribution, which implies that

$$E(X) = V(X) = m. \quad (6)$$

It can be shown that, taking all junctions of the “same” type, the number of accidents X has a negative binomial distribution. In order to determine $E(X)$ and $V(X)$, it is possible to make use of the fact that X here is a double stochastic variable. Thus it follows that:

$$E(X) = E[E(X|m)] = E(m) = pred \quad (7)$$

and that

$$\begin{aligned} V(X) &= E[V(X|m)] + V[E(X|m)] \\ &= E(m) + V(m) \\ &= pred + \frac{pred}{p}. \end{aligned} \quad (8)$$

Fig. 1. Frequency function for a gamma-distributed stochastic variable m with expected value “ $pred.$ ”

*The frequency function for the gamma distribution is

$$f(m) = \frac{p(pm)^{q-1}e^{-pm}}{\Gamma(q)}, m > 0$$

It may be noted that this variance is greater than for the Poisson distribution. There is equality only when all the junctions of the "same" type have the same m , i.e. when $V(m) = \frac{pred}{p} = 0$.

By using the double stochastic distribution, it is possible to rewrite the eqns (1) and (2) and obtain the following alternative EB estimate for a chosen junction with a predicted number of accidents equal to " $pred$ " and an observed number of accidents equal to x :

$$\hat{m} = x + k(pred - x) \quad (9)$$

where

$$k = \frac{pred}{pred + \frac{pred}{p}} = \frac{1}{1 + \frac{1}{p}} = \frac{1}{1 + \frac{pred}{q}} = \frac{1}{1 + a \cdot pred} \quad (10)$$

Alternatively, (9) can be written as

$$\hat{m} = pred + k'(x - pred) \quad (11)$$

where

$$k' = 1 - k \quad \text{i.e.} \quad k' = \frac{a \cdot pred}{1 + a \cdot pred} \quad (12)$$

Since $pred \geq 0$ both k and k' will be in the range from 0 to 1. When $pred$ and/or $a \rightarrow 0$ then $k' \rightarrow 0$ ($k \rightarrow 1$) and $\hat{m} \rightarrow pred$ i.e. the predicted value. When instead $pred$ and/or $a \rightarrow \infty$ then $k' \rightarrow 1$ ($k \rightarrow 0$) and $\hat{m} \rightarrow x$ i.e. the observed value.

The magnitude of the correction factor k' (alternatively k) will depend both on " $pred$ " and on the precision of the prediction model. The better the precision of the prediction model, the smaller the magnitude of a will be.

The diagram in Fig. 2 illustrates how k' varies with " $pred$ " for various values of a .

EMPIRICAL COMPARISON OF DIFFERENT METHODS OF CORRECTING FOR THE EFFECT DUE TO REGRESSION-TO-THE-MEAN

Below are described two empirical methods of estimating the expected number of accidents m and thereby correcting for the regression effect in nonexperimental before-and-after studies.

The methods compared are the commonly used EB method [eqn (2)] and the variant of the EB method described in this report [eqn (11)] which uses prediction models. In the case of the variant of the EB method different values of a have been used (0.04, 0.10, 0.20, 0.30, and 1).

The material used in Table 1 consists of accidents at 1,901 3-way junctions on rural roads. The before period comprises the years 1977–1981 and the after period the years 1982–1983.

It should be noted that in the case of junctions with a certain number of accidents (e.g. 0) during the before period, the same estimated m is obtained for each and every one of these junctions with the conventional EB method while the estimates differ when the variant of the method is used. This is because " $pred$ " varies from junction to junction owing to different traffic flows.*

* $pred = 0.000069 \cdot (I_p + I_s)^{1.0957} \cdot (I_s/I_p + I_s)^{0.3052}$ where I_p and I_s are the numbers of incoming vehicles per day (and night) from the primary and secondary road respectively.

Fig. 2. Correction factor k' as a function of "pred" for different values of a .

Table 1 shows that the EB variant with $a = 0.04$ gives estimates of m very close to the predicted values. The EB variant with $a = 1$, however, gives estimates of m that are rather close to the observed values for the before period, 1977–81. The EB variant with the other values of a and the conventional EB method give results fairly similar to each other and agreement with the "answer," i.e. the annual number of accidents per junction for the after period 1982–83, is also good.

It is not therefore possible on the basis of the above results to demonstrate any better estimates with the EB variant than the conventional EB method, even though in the former case predicted values are used for each individual junction. Nor has it been possible to demonstrate any significant difference between the two methods in terms of the dispersion of the estimates. Additional correlation and regression analyses are shown in Appendix A.

If "pred" were set equal to $\bar{x}$ throughout for the EB variant, the estimates would be very poor for values of a between 0.04–0.30. For $a = 1$, however, the estimates would be almost the same as for the conventional EB method and therefore good.

In Table 2, the comparison between the different methods is limited to the subset of 100 junctions with the lowest predicted numbers of accidents. When using the conventional EB method, $\bar{x}$ and s^2 have been based on the whole population. This means that the m estimates for this method will be the same as in Table 1.

According to Table 2, the variant of the EB method (with the exception of $a = 1$) gives very good estimates $\hat{m}$. On the other hand, the conventional EB method are not at all good since this method overestimates m greatly. This is not surprising since $\bar{x}$ and s^2 here have been calculated for the entire population of 1,901 junctions and that a correction has consequently been made for each observation unit in the direction of $\bar{x}$. In order for the conventional EB method to have given good estimates of m , it would

Table 1. Estimated values of m , all junctions

No. of junct.	No. of acc. 1977-81	Annual no. of acc. per junct. 1977-81	Pred. annual no. of acc. per junct. 1977-81*	$\hat{m}$ for annual no. of accidents per junction 1977-81				Annual no. of acc. per junct. 1982-83†	
				EB conv.†	EB $a = 0.04$	EB $a = 0.10$	EB $a = 0.20$	EB $a = 0.30$	EB $a = 1$
727	0	0	0.18	0.11	0.17	0.16	0.15	0.14	0.09
559	1	0.20	0.23	0.23	0.22	0.22	0.22	0.21	0.20
277	2	0.40	0.28	0.35	0.28	0.28	0.29	0.30	0.33
155	3	0.60	0.34	0.47	0.35	0.37	0.39	0.41	0.48
68	4	0.80	0.40	0.59	0.42	0.45	0.49	0.52	0.64
40	5	1.00	0.46	0.71	0.50	0.54	0.61	0.65	0.81
75	≥6	1.57	0.64	1.05	0.74	0.86	0.99	1.08	1.33

* Also applies to 1982-83.

† For the years 1977-81 $\bar{x} = 1.39$ and $s^2 = 3.46$.

‡ Multiplied by 1.10 to correct for the general accident trend.

Table 2. Estimated values of m , limited to junctions with the lowest predicted numbers of accidents

No. of junct.	No. of acc. 1977-81	Annual no. of acc. per junction 1977-81	Pred. annual no. of accidents per junct. 1977-81*	$\hat{m}$ for annual no. of acc. per junction 1977-81				Annual no. of acc. per junction 1982-83†
				EB conv.†	EB $a = 0.04$	EB $a = 0.10$	EB $a = 0.20$	EB $a = 1$
73	0	0	0.08	0.11	0.08	0.08	0.08	0.08
20	1	0.20	0.08	0.23	0.09	0.09	0.09	0.08
7	2	0.40	0.07	0.35	0.08	0.08	0.09	0.08

* Also applies to 1982-83.

† Using $\bar{x} = 1.39$ and $s^2 = 3.46$.

‡ Multiplied by 1.10 to correct for the general accident trend.

have been necessary first to have calculated and utilized $\bar{x}$ and s^2 only for the 100 junctions with the lowest predicted numbers of accidents.

If the junctions with the highest predicted numbers of accidents had been chosen instead, the variant of the EB method would also have given good estimates of m while the conventional EB method (with $\bar{x}$ and s^2 based on the whole population) would of course have underestimated m .

The example shown in Table 2 is of course extreme. But even in practice it has been found that serious problems may occur in defining a suitable subpopulation. In addition, the estimates $\bar{x}$ and s^2 become more and more uncertain the smaller the population on which the calculations are based.

EXAMPLE OF PRACTICAL APPLICATION

The following example is from an actual case. A certain type of signal control has been introduced at 6 junctions and the effect of this countermeasure on the number of accidents is to be determined. It may be supposed that a large number of accidents is one of the reasons for introducing signal control at the junctions. In this case, this would imply that the effect of the countermeasure would probably be overestimated in a before-and-after study unless the regression effect is eliminated.

It has been difficult for several reasons to apply the conventional EB method. It has not been possible to define a suitable population of junctions. Furthermore, it would have been laborious to collect accident data for the population. The situation is also complicated by the fact that the before and after periods differ from junction to junction. Because of these factors, only the variant of the EB method (for different values of a) has been tested.

On the average, the observed number of accidents during the before periods is greater than predicted. Note, however, that for one junction the observed and predicted numbers of accidents are approximately the same and that for two of the junctions the observed number of accidents is in fact less than predicted.

Since this is a matter of junctions with fairly high flows and thus also fairly high accident levels, the regression effect ought not to be especially pronounced. According to the model curve used earlier [Brüde *et al.*, 1985] and which is reproduced in Appendix B the size of the regression effect in this case should be $\geq 5\%$.

From Table 3, estimates are obtained of the regression effect by comparing " $\hat{m}$ before" (reckoned over all 6 junctions) with the observed number of accidents during the before periods. The regression effect will then be anything from $11\% = (1 - 68.5/77) \cdot 100\%$ when $a = 0.04$ to 2% when $a = 1$.

Estimates of the countermeasure effect are obtained by comparing the observed number of accidents during the after periods for all 6 junctions with " $\hat{m}$ after if no countermeasure." The estimated countermeasure effect then varies between $1\% = (1 - 57/57.4) \cdot 100\%$ when $a = 0.04$ to 17% when $a = 1$.

If the judgement is made that a in the case of the variant of the EB method should here be between 0.04 and 0.30, the estimated countermeasure effect could then vary from effectively 0% to almost 15%.

DISCUSSION AND CONCLUSIONS

When it is possible to define a population of objects (e.g. junctions) that is not too small and that corresponds well to the population currently of interest, and when it is possible also to collect accident data for this population, then the conventional EB method is satisfactory. Although the variant of the EB method uses predicted numbers of accidents for each individual object (in this case junctions) it has not been possible to demonstrate that the average estimates of the expected number of accidents are better than those obtained with the conventional method.

Table 3. Estimate of the effects due to regression-to-the-mean and due to countermeasure using the EB variant (varying values of a) for 6 junctions where a countermeasure has been implemented

Junct. no.	Obs. no. of acc. bef.	Pred. no. of acc. bef.	$\hat{m}$ before					Ratio of before period to after period	$\hat{m}$ after if no countermeasure used					Obs. no. of acc. after
			EB $a = 0.04$	EB $a = 0.10$	EB $a = 0.20$	EB $a = 0.30$	EB $a = 1$		EB $a = 0.04$	EB $a = 0.10$	EB $a = 0.20$	EB $a = 0.30$	EB $a = 1$	
1	11	8.8	9.4	9.8	10.2	10.4	10.8	1.77	16.6	17.3	18.1	18.4	19.1	24
2	12	12.4	12.3	12.2	12.1	12.1	12.0	0.85	10.5	10.4	10.3	10.3	10.2	8
3	18	6.7	9.1	11.2	13.2	14.2	16.5	1.20	10.9	13.4	15.8	17.0	19.8	8
4	16	10.2	11.9	13.1	14.1	14.6	15.5	0.66	7.9	8.6	9.3	9.7	10.2	8
5	5	10.4	8.8	7.6	6.8	6.3	5.5	0.40	3.5	3.0	2.7	2.5	2.2	4
6	15	18.4	17.0	16.2	15.7	15.5	15.2	0.47	8.0	7.6	7.4	7.3	7.1	5
Total	77	66.9	68.5	70.1	72.1	73.1	75.5		57.4	60.3	63.6	65.2	68.6	57
			Estimated regression effect (%)						Estimated countermeasure effect (%)					
			11	9	6	5	2		1	5	10	13	17	

However, if it is only a certain subpopulation that is of interest, while the only information available concerns the whole population when using the EB method, the estimates obtained will be poorer than for the variant of the EB method.

In certain cases, it is more or less impossible to obtain the information required for the conventional EB method. Here the variant of the EB method may be the only conceivable method.

In the case of the variant of the EB method, it has still not been ascertained how the precision of the prediction model should generally be taken into account. This means that different assumptions concerning precision lead to varying results, e.g. when attempting to estimate the effect of a certain countermeasure in a nonexperimental before-and-after study. Further research is necessary in order to clarify this problem.

REFERENCES

- Abbess C., Jarret D. and Wright C. C., Accidents at blackspots: Estimating the effectiveness of remedial treatment, with special reference to the "regression-to-mean" effect. *Traffic Eng. Control* **22**:535-542, 1981.
- Brüde U. and Larsson J., The "regression-to-mean" effect. Some empirical examples concerning accidents at road junctions. VTI Report 240, 1982.
- Brüde U. and Larsson J., Countermeasures taken at junctions as part of the regional road authorities' traffic safety program. Effects due to the regression-to-the-mean and to the countermeasures respectively. VTI Report 292, 1985.
- Danielsson S., A comparison of two methods for estimating the effect of a countermeasure in the presence of regression effects. *Accid. Anal. Prev.* **18**:13-23, 1986.
- Hauer E., On the estimation of the expected number of accidents. *Accid. Anal. Prev.* **18**:1-12, 1986.
- Junghard O. and Danielsson S., Accident rates and compound probability distribution. VTI Bulletin 476, 1986.

APPENDIX A

CORRELATION AND REGRESSION ANALYSIS

- N = number of observations = 1,901 junctions
- A8283 = annual no. of accidents 1982-83 multiplied by 1.10
- A7781 = annual no. of accidents 1977-81
- P = predicted annual no. of accidents
- EB = annual no. of accidents 1977-81 according to the conventional EB method
- EB1 = annual no. of accidents 1977-81 when $a = 0.04$
- EB2 = annual no. of accidents 1977-81 when $a = 0.10$
- EB3 = annual no. of accidents 1977-81 when $a = 0.20$
- EB4 = annual no. of accidents 1977-81 when $a = 0.30$
- EB5 = annual no. of accidents 1977-81 when $a = 1$

Correlation matrix

	A8283	A7781	P	EB	EB1	EB2	EB3	EB4	EB5
A8283	1.00	0.42	0.46	0.42	0.49	0.50	0.50	0.49	0.46
A7781	0.42	1.00	0.62	1.00	0.72	0.81	0.87	0.91	0.97
P	0.46	0.62	1.00	0.62	0.99	0.95	0.90	0.86	0.74
EB	0.42	1.00	0.62	1.00	0.72	0.81	0.87	0.91	0.97
EB1	0.49	0.72	0.99	0.72	1.00	0.99	0.96	0.93	0.83
EB2	0.50	0.81	0.95	0.81	0.99	1.00	0.99	0.97	0.91
EB3	0.50	0.87	0.90	0.87	0.96	0.99	1.00	0.99	0.96
EB4	0.49	0.91	0.86	0.91	0.93	0.97	0.99	1.00	0.98
EB5	0.46	0.97	0.74	0.97	0.83	0.91	0.96	0.98	1.00

Linear regression analysis ($\hat{y} = a + bx$)

y	x	a	b	R^2
A8283	A7781	0.11	0.55	0.18
A8283	P	-0.07	1.34	0.22
A8283	EB	0.01	0.91	0.18
A8283	EB1	-0.07	1.32	0.24
A8283	EB2	-0.06	1.24	0.25
A8283	EB3	-0.03	1.11	0.25
A8283	EB4	-0.01	1.02	0.24
A8283	EB5	0.05	0.77	0.21
EB1	P	-0.01	1.05	0.97
EB5	A7781	0.06	0.76	0.95
EB1	EB	0.10	0.58	0.52
EB2	EB	0.06	0.70	0.65
EB3	EB	0.03	0.85	0.77
EB4	EB	0.002	0.95	0.83
EB5	EB	-0.08	1.27	0.95

APPENDIX B

“Normal regression effect” as a function of the mean number of accidents in the population from which the biased sample has been taken.