Statistica

Netherlands School of Economics
ECONOMETRIC INSTITUTE

Report 7215

GENERAL LEAST-SQUARES ESTIMATION OF LINEAR RELATIONSHIPS WITH POISSON-DISTRIBUTED DEPENDENT VARIABLES RELATING TO HETEROGENEOUS OBSERVATIONS

by Bob van der Laan

July 21, 1972

Preliminary and Confidential

GENERAL LEAST-SQUARES ESTIMATION OF LINEAR RELATIONSHIPS WITH POISSON-DISTRIBUTED DEPENDENT VARIABLES RELATING TO HETEROGENEOUS OBSERVATIONS

by Bob van der Laan[1]

Contents

1. INTRODUCTION AND SUMMARY

General theories or methods cannot often be used in real-life problems, because the assumptions made for obtaining elegant results are unrealistic in specific problems. General theories and methods are not as generally applicable as the word suggests.

This paper deals with a number of problems encountered in the application of the general linear model

$$(1.1)\quad y = X\beta + u$$

to special cases. In (1.1) y denotes a vector of n values assumed by the dependent variable, X is a matrix of order $n \times k$ of non-stochastic values assumed by the explanatory variables, β is a vector of k unknown parameters, and u is a vector of unknown random variables (the disturbances)

[1] The author wishes to thank Prof.Dr. W.H.Somermeyer and Prof. Dr. A.P.J. Abrahamse for their valuable help in preparing this paper. The paper is presented at the European Meeting of the Econometric Society, Budapest, September 1972.

The first problem reads as follows. The vector u is usually supposed to be normally distributed. This assumption implies that the vector y in (1.1) is also normally distributed. Sometimes, however, there are a priori reasons to suppose that y is not normally distributed: for example, if y can assume integer values only. If the normality assumption is dropped, many properties, derived for the case where u is normally distributed, are no longer valid.

Second, assuming that the dependent variable is Poisson-distributed, the small number of classes becomes an acute problem. For instance, if the Poisson parameter is 0.1, the probability of 2 or more "successes" is only 0.0047, hence, in practice, the values of y_i will be confined to 0 or 1. Even if $\lambda = 3.0$, the probability of 8 or more "successes" is only 0.0119; this means that, in fact, y_i is restricted to the values 0,1, ..., 7. Statistical studies have shown that the relative number of motorists involved in 2 or more accidents in a year is generally very small. In such cases the number of classes to be distinguished is small; consequently it is difficult to decide whether a hypothesis about the linear dependence of some variables should be accepted or rejected.

Third, one generally assumes that the observed values of y_i relate to the same reference set. This assumption does not always hold good in specific problems. For example, some observed values relate to an entire year, while others relate to a part of a year only. If one wishes to make use of all available information, i.e. including data related to only a part of a year, the question arises as to how it should be incorporated into the analysis of the problem.

The problems considered in this paper presented themselves during the examination of the number of accidents in which a motorist is involved in a given time period and the resulting amount of damage; the effects of certain factors on the number of accidents and on the amount of damage are studied simultaneously. One often assumes that the number of accidents in which a motorist is involved in some time period is Poisson-distributed, with the

parameter λ very small,
viz. between 0.1 to 0.5. The hypothesis is confirmed by a sizeable amount of information concerning car accidents. The usual regression procedure applied to such a number of observations is rather time-consuming, in particular, if an iterative estimation procedure is used.

Therefore, we try to apply a method of grouping observations. Prais and Aitchison (1954) "give a rigorous treatment of certain problems that arise in applying regression techniques to grouped observations" (cf. p. 1). Because of the different structure of our specific problem, our approach deviates somewhat from the procedure adopted by Prais and Aitchison. In Section 2 we construct a linear model, based on grouped observations, in which the dependent variable is Poisson-distributed. In Section 3 we apply this model to observations related to number of accidents and amount of damage. In Section 4 we present empirical applications of the models set out in Section 3. Finally, we give some concluding remarks in Section 5.

2. THE MODEL

Consider a sample of n observations on variables $(y_i;x_{i2}, \ldots, x_{ik})$, with y_i the i^{th} observation of a random variable Y_i, denoting the number of events occurring in an interval $(0,N_i)$ and $x_{i2}, \ldots, x_{ik}$ the corresponding values of the (assumedly) non-stochastic explanatory variables. The interval $(0,N_i)$ may be specified as a time period, a linear measure, an area, a volume, etcetera; for the sake of simplicity, however, we assume in this section that the interval is a time period of N_i months all of which are supposed to be of equal length. In Section 3 we also consider the interval as a distance.

We define the random variable Z as the number of events occurring in an interval of one month, with Z

Poisson-distributed with parameter λ_i. The variable Y_i may then be considered as being the sum of N_i Z-variables. If we assume that the variables Z_l, $l=1, \ldots, N_i$ are independently distributed, Y_i is Poisson-distributed with parameter $N_i\lambda_i$.

We assume that λ_i is some function of the variables $x_2, \ldots, x_k$. In general, this will be a non-linear function, but it can often be approximated by a linear one within a set Ω of values of $x_2, \ldots, x_k$. If λ_i is a linear function (or an approximation thereof) of the variables $x_2, \ldots, x_k$ for $(x_2, \ldots, x_k)\epsilon\Omega$, one gets

$$(2.1)\quad \lambda_i = \beta_1 + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} \text{ for } (x_{i2}, \ldots, x_{ik})\epsilon\Omega \qquad (i = 1, \ldots, n)$$

Consequently, the following regression model is obtained:

$$(2.2)\quad y_i = N_i(\beta_1 + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}) + u_i \quad \text{for } (x_{i2}, \ldots, x_{ik})\epsilon\Omega \quad (i = 1, \ldots, n)$$

in which y_i relates to a time period of N_i months, and u_i is the disturbance term with expectation equal to zero and variance equal to $N_i(\beta_1 + \Sigma\beta_j x_{ij})$. We assume that the disturbances are stochastically independent.

Let

$$(2.3)\quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix},\ u = \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix},\ \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix},\ X = \begin{bmatrix} 1 & x_{12} & \ldots & x_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n2} & \ldots & x_{nk} \end{bmatrix}$$

$$N = \begin{bmatrix} N_1 & & 0 \\ & \ddots & \\ 0 & & N_n \end{bmatrix}, \text{ and } \Lambda = \begin{bmatrix} \beta_1 + \sum_{j=2}^{k} \beta_j x_{1j} & & 0 \\ & \ddots & \\ 0 & & \beta_1 + \sum_{j=2}^{k} \beta_j x_{nj} \end{bmatrix} = \operatorname{diag}\{X\beta\}$$

Using (2.3) we can write (2.2) as

$$(2.4)\quad y = NX\beta + u$$

with $E(u) = 0$ and $E(uu') = N\Lambda$.

The generalized least-squares (G.L.S.) estimator of β is:

$$(2.5)\quad \hat{\beta} = (X'N\Lambda^{-1}X)^{-1}X'\Lambda^{-1}y$$

According to the well-known Gauss-Markov theorem, the G.L.S. estimator is a "best" estimator, in the sense that it minimizes the mathematical expectation of any positive semi-definite quadratic form in the sampling errors, within the class of linear unbiased estimators. The variance-covariance matrix $N\Lambda$ is unknown, however, as it depends on the vector of unknown parameters β. Hence, the Gauss-Markov theorem cannot be applied here. Since $\hat{\beta}$ cannot be determined directly, we have to apply an iterative estimation procedure.

Defining $\hat{u} = y - NX\hat{\beta}$, we can estimate the i^{th} diagonal element of $N\Lambda$, d_i^2, in the n^{th} cycle of the iteration by means of $\hat{u}_i^2$ obtained in the $(n-1)^{th}$ cycle. Independent of each other, C.R. Rao (1970) and V. Chew (1970) derived similar estimators for a diagonal variance-covariance matrix, named Minimum Norm Quadratic Unbiased Estimation (MINQUE) by Rao. This method does not, however, preclude negative variance estimates. In such cases Chew suggests that "negative estimates are replaced by zeroes or else quadratic procedures are used in the least-squares solution to constrain the estimates to be non-negative" (p. 175). Chew does not discuss the consequences of these two solutions of the problem of possibly negative estimates. Moreover, he does not show in which cases his estimation procedure leads to negative estimates. Applying this method to six sets of values of y, N, and

X, we got negative estimates of one or more d_i^2 in four cases. Therefore, this method seems to be unsuitable for solving our problems.

We prefer to apply the estimation procedure described by Prais and Aitchison (1954 p. 18) in order to estimate β and the variance-covariance matrix. For the special case where y is Poisson-distributed, Λ and β can well be estimated iteratively by means of a method described by Prais (1953). Adopting the unit matrix as the estimate of Λ in the first cycle of the iteration, the estimators of Λ and β in the r^{th} cycle are specified by

$$(2.6)\quad \hat{\Lambda}_{(r)} = \text{diag}\ \{X\hat{\beta}_{(r-1)}\}$$

and

$$(2.7)\quad \hat{\beta}_{(r)} = (X'N\hat{\Lambda}_{(r)}^{-1}X)^{-1}\ X'\hat{\Lambda}_{(r)}^{-1}y$$

respectively. The estimation (2.7) is not unbiased, and its variance-covariance matrix is unknown (cf., e.g. Weber (1971)). Jorgenson states $\hat{\beta}_{(r)}$ is best asymptotically normal, and proves that the iterative procedure converges, provided that $\hat{\Lambda}_{(r)}$ and $(X'N\hat{\Lambda}^{-1}X)^{-1}$ are positive definite per all r.

Some remarks must be made with respect to the above-mentioned estimation procedure.
First, the Poisson variable can assume non-negative (integer) values only. If we estimate $\hat{\beta}$ according to (2.7), the possibility of negative values of the elements of $NX\hat{\beta}$ is not ruled out. In order to exclude negative values we must estimate β subject to the constraint $NX\hat{\beta} \geq 0$ for all $(x_2, \ldots, x_k) \epsilon \Omega$. The linear regression problem then becomes a quadratic programming problem. This implies an increase in computation time.
Second, if λ is small, y will assume a few values only. In such a case, the G.L.S. estimation procedure is of little or no use. Instead we could adopt another esti-

mation method, a method, for instance, such as Wald's device of fitting straight lines to sub-sets of variables ranked according to the value of one or more of the explanatory variables. However, Wald's method too does not prevent the possibility of elements of $NX\beta$ assuming negative values. Alternatively, we could apply estimation methods based on discriminant analysis, probit analysis or logit analysis. An objection against the application of these models is that the dependent variable is measured on a nominal scale, while the dependent variable in model (2.4) is measured on the ratio scale. Therefore, applying such methods would mean wasting information. Moreover, Wald's method and the methods based on discriminant analysis, probit analysis or logit analysis do not take into account that the observed values of y_i may relate to different reference sets.
Third, if the number of observations is large, an iterative estimation procedure is time-consuming.

In this paper we tackle the above-mentioned difficulties by grouping the observations in a number of groups; this approach can be seen as an extension of model (2.4). Methods of grouping are described, for instance, by Prais and Aitchison (1954), and Cramer (1964). We classify the sample elements on the basis of the values of the explanatory variables. For the sake of simplicity of presentation we assume a single explanatory variable only. Hence the regression model (2.2) is reduced to:

$$(2.8)\quad y_i = N_i(\beta_1 + \beta_2 x_i) + u_i \quad (i = 1, \ldots, n)$$

for x_i in some interval (x', x'').
Suppose that the sample of observations can be divided into G homogeneous groups S_g $(g = 1, \ldots, G)$ according to the values of x_i. Summation of (2.8) over all $i \epsilon S_g$ yields:

$$(2.9)\quad \sum_{i\epsilon S_g} y_i = \sum_{i\epsilon S_g} N_i(\beta_1 + \beta_2 x_i) + \sum_{i\epsilon S_g} u_i \quad (g = 1, \ldots, G)$$

Let $N_g^* = \sum_{i \epsilon S_g} N_i$ = the total number of months corresponding to group S_g;

$y_g = \sum_{i \epsilon S_g} y_i$ = the total number of events occurring within group S_g;

$\bar{x}_g = \frac{1}{N_g^*} \sum_{i \epsilon S_g} N_i x_i$ = the weighted mean of the values of the explanatory variable in group S_g;

and $u_g^* = \sum_{i \epsilon S_g} u_i$, hence $E(u_g^*) = 0$ and $\text{var}(u_g^*) = N_g^*(\beta_1 + \beta_2 \bar{x}_g)$

As a result, (2.9) can be rewritten as

$$(2.10)\quad y_g^* = N_g^*(\beta_1 + \beta_2 \bar{x}_g) + u_g^* \qquad (g = 1, \ldots, G).$$

Writing:

$$(2.11)\quad \begin{cases} y^* = \begin{bmatrix} y_1^* \\ \vdots \\ y_G^* \end{bmatrix}, \; N^* = \begin{bmatrix} N_1^* & & 0 \\ & \ddots & \\ 0 & & N_G^* \end{bmatrix}, \; \bar{X} = \begin{bmatrix} 1 & \bar{x}_1 \\ \vdots & \vdots \\ 1 & \bar{x}_G \end{bmatrix}, \\ \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, \; u = \begin{bmatrix} u_1^* \\ \vdots \\ u_G^* \end{bmatrix}, \text{ and } \bar{\Lambda} = \begin{bmatrix} \beta_1 + \beta_2 \bar{x}_1 & & 0 \\ & \ddots & \\ 0 & & \beta_1 + \beta_2 \bar{x}_G \end{bmatrix}, \end{cases}$$

(2.10) reads in matrix notation as

$$(2.12)\quad y^* = N^* X \beta + u^*$$

with $E(u^*) = 0$ and $E(u^* u^{*\prime}) = N^* \bar{\Lambda}$. The G.L.S. estimator of β equals

$$(2.13)\quad \hat{\beta} = (\bar{X}' N^* \Lambda^{-1} \bar{X})^{-1} \bar{X}' \Lambda^{-1} y^*$$

which can be approximated iteratively, similar to (2.7).

The possibility of negative values of the elements of $N^* \bar{X} \hat{\beta}$ is not ruled out, but it will be less probable than the occurrence of negative values of $NX\hat{\beta}$ in model

(2.4): while in model (2.4), y_i assumes the values 0, 1, or 2 only, we now have y_g^*, which can assume many more different values. This makes possible the application of the G.L.S. estimation method. Since G is much smaller than n, the iterative procedure for estimating (2.13) takes much less time than the one required for estimating (2.5). On the other hand, applying this approach means, of course, a loss of information. By choosing G sufficiently large, however, the latter disadvantage will be compensated for the above-mentioned advantages.

The method of grouping of the observations can be generalized to more than one explanatory variables in a simple manner. The qualitative and quantitative criteria for grouping depend on the number of observations, the number of classes per explanatory variable and the number of explanatory variables.

3. THEORETICAL APPLICATIONS IN THE FIELD OF MOTOR CAR ACCIDENTS

3.1. Number of Accidents

Let the number of accidents in which a motorist is involved in one month be Poisson-distributed with parameter λ, while λ is assumed to be a linear function of k-1 risk factors $x_2, \ldots, x_k$:

$$(3.1)\quad \lambda = \beta_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

In general, however, the observations do not relate to time intervals of equal length. Still, observation y_i, related to N_i months, can be assumed to be Poisson-distributed with parameter $N_i\lambda_i$. If this assumption is satisfied, we get the model (2.2), or - in matrix notation - (2.4).

If we group the observations according to a single explanatory variable, we get model (2.10), with y_g^* denoting the total number of accidents in which motorists in group S_g were involved, and N_g^* denoting the total

number of months these motorists drove.

A motorist, involved in a number of accidents in a given time period, has driven a number of kilometres in that period. It is only reasonable to assume that the more kilometres a motorist drives, the greater probability of his becoming involved in one or more accidents. One of the explanatory factors in (3.1) will therefore be the number of kilometres driven during one month. We can, however, consider the problem in another way. A motorist i, who drives N_i kilometres, can be assumed to perform N_i random experiments with a number of accidents as outcomes. Let y_i denote this number of accidents. As has been shown by Van der Laan (1971), y_i is Poisson-distributed with parameter $N_i\lambda_i$, with λ_i a constant, given a number of assumptions. If λ_i depends linearly on t - 1 explanatory factors

$$(3.2)\quad \lambda_i = \beta_1 + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$$

we get the following regression model

$$(3.3)\quad y_i = N_i(\beta_1 + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}) + u_i \quad (i = 1, \ldots, n)$$

In this case too, we can group the observations. As a result we get a regression model as described in (2.10), with y_g^* denoting the total number of accidents in which the motorists in group S_g were involved, as before, while N_g^* now represents the total number of kilometres these motorists drive.

3.2. Amount of Damage

In this subsection we apply the method of grouping outlined in Section 2 to a model for the amount of damage incurred. In the literature on accident statistics the assumption is often made that the amount of damage caused by motorists is lognormally distributed. Van der Laan and Boermans (1970) also assumed a lognormal distribution before elimination of systematic effects on the basis of

empirical evidence. Accepting such a distribution, the use of a multiplicative model seems indicated. Therefore, we posit the following relationship [2]

$$(3.4)\quad z_i = B e^{\beta_2 x_i} v_i \qquad (i = 1, \ldots, n)$$

where n is the number of accidents in a given time period with amount of damage exceeding zero; z_i denotes the amount of damage implied by the i^{th} accident; x_i is a non-stochastic value of the explanatory variable; B and β_2 are constants; and v_i is a disturbance term which is lognormally $(1,\sigma_i^2)$ distributed. Furthermore, we assume that the disturbances are stochastically independent. Taking the natural logarithm on both sides of the equality sign in (3.4), we get

$$(3.5)\quad \ln z_i = \ln B + \beta_2 x_i + \ln v_i \qquad (i = 1, \ldots, n)$$

Putting $y_i = \ln z_i$, $\beta_1 = \ln B$ and $u_i = \ln v_i$, (3.5) can be written as

$$(3.6)\quad y_i = \beta_1 + \beta_2 x_i + u_i \qquad (i = 1, \ldots, n)$$

with u_i $N(0,\sigma_i^2)$ distributed.

According to the method of grouping presented in Section 2, we classify the set of observations in G groups S_g. Summation of (3.6) over all $i\epsilon S_g$ yields:

$$(3.7)\quad y_g^* = N_g^*(\beta_1 + \beta_2 \bar{x}_g) + u_g^* \qquad (g = 1, \ldots, G)$$

with N_g^* = the total number of accidents in group S_g,

$\bar{x}_g = \frac{1}{N_g^*} \sum_{i\epsilon S_g} x_i$, $y_g = \sum_{i\epsilon S_g} y_i$, while $u_g^* = \sum_{i\epsilon S_g} u_i$

is normally distributed with zero mean and variance equal to $d_g^2 = \sum_{i\epsilon S_g} \sigma_i^2$. Defining

---

[2] For the sake of simplicity we again consider the case of one explanatory variable only.

$$(3.8)\quad D = \begin{bmatrix} d_1^2 & & 0 \\ & \ddots & \\ 0 & & d_G^2 \end{bmatrix}$$

and using the definitions of (2.11), (3.7) reads in matrix notation:

$$(3.9)\quad y^* = N^*\bar{X}\beta + u^*$$

with u^* $N(0,D)$ distributed. This model is similar to the one presented in Section 2, except for the assumption concerning the distribution of u^*. The G.L.S. estimator of β is expressed by

$$(3.10)\quad \hat{\beta} = (\bar{X}'N^*D^{-1}N^*\bar{X})^{-1}\bar{X}'N^*D^{-1}y^*$$

Since the variance-covariance matrix D is unknown, we can estimate β and D by means of an iterative procedure. If we assume homoskedasticity, implying $\sigma_i^2 = \sigma^2$ for all i, we get $D = \sigma^2 N^*$. Hence, the estimators of β and σ^2 are simply:

$$(3.11)\quad \hat{\beta} = (\bar{X}'N^*\bar{X})^{-1}\bar{X}'y^*$$

and

$$(3.12)\quad s^2 = \frac{1}{G-2}\,\hat{u}^{*\prime}N^{*-1}\hat{u}^*$$

respectively.

4. EMPIRICAL APPLICATIONS

The approach adopted in this paper is applied to the models of Section 3, with the age of the car as the only explanatory variable, and applying the number of accidents per kilometer, the number of accidents per month, and the amount of damage as the successive dependent variables, respectively. Of course, the age of the car is not the only factor which possibly

influences the number of accidents a motorist is involved or the amounts of the damages. Other factors may be, for example, the age and sex of the motorist, the area where he usually drives, his driving experience, etc.

We distinguish between two types of insurances, viz.: a) third-party insurance, which is an insurance against the risk of causing damage to others, and b) casco insurance (or insurance on hull and appurtenances) which insures one's own car against damage or loss. The data is confined to accidents reported to the company and for which indemnification has been paid.[3] The age of the car is defined as the difference between the year in question and the year in which the car was manufactured.

We use data from a sample of Dutch insurance policies of cars and station-cars relating to the years 1963, 1964 and 1965. The sample elements are grouped according to the age of the car. The definition of the age gives rise to errors of measurement. The ensueing distortions will be partly cancelled out, however, by the grouping. The average numbers of kilometres which the motorists drive per year was not incorporated in the sample of the insurance policies, but have been added later on. They have been taken from tables drawn up by the Netherlands Central Bureau of Statistics (cf. C.B.S. (1965) and (1967)). An extra advantage of the grouping procedure is that for grouped observations we may use data about the average number of kilometers driven per year by groups of motorists which are known, instead of individual data, which are in general unknown to insurance companies. Table 1 presents the grouped data used. These imperfections in the data must be taken into account in interpreting the results. The applications in this section only serve as examples.

First, we assume that the number of accidents per kilometre is linearly dependent on the age of the car.

---

[3] We have to distinguish between the number of accidents made in a given time-period and the number of accidents reported in the same time-period. The second number of accidents is smaller than the first, for not all accidents are reported to the company (cf. Van der Laan (1971).

Table 1. Distribution of the absolute and relative numbers of accidents and the average amount of damage, per age group of the car.

Type of insurance	Age of the car	Average number of kilometres driven per year*	Total number of car years	Total number of accidents	Average number of accidents per 18,000 km/year	Average number of accidents per car year	Average amount of damage in Dfl.
	1	2	3	4	5	6	7
Third-party	0	23000	627	84	0.1048	0.1340	390
	1	21400	1168	131	0.0942	0.1122	452
	2	18800	870	91	0.1001	0.1046	427
	3	17300	663	73	0.1146	0.1102	367
	4	16000	574	62	0.1214	0.1080	401
	5	14400	503	46	0.1146	0.0915	350
	6	13300	433	52	0.1628	0.1201	551
	7	12500	369	42	0.1634	0.1137	404
	8	11900	382	40	0.1589	0.1047	341
	9	11400	326	38	0.1838	0.1164	239
	10	10900	265	38	0.2372	0.1434	212
	11	9700	211	25	0.2208	0.1187	603
	12	8600	167	16	0.2008	0.0959	681
	13	7600	118	12	0.2422	0.1016	534
	$\geq$ 14	6800	150	19	0.3367	0.1270	692
casco	0	23000	548	197	0.2814	0.3598	379
	1	21500	970	387	0.3342	0.3989	496
	2	18900	632	211	0.3186	0.3339	499
	3	17300	338	93	0.2858	0.2750	388
	4	16200	187	64	0.3814	0.3424	482
	5	14400	107	33	0.3858	0.3096	347
	6	13300	57	22	0.5231	0.3877	607
	$\geq$ 7	12000	43	10	0.3517	0.2339	350

* cf. C.B.S. (1965) and (1967).

We use the regression model

$$(4.1)\quad y^* = N^*\bar{X} + u^*$$

with $E(u^*) = 0$ and $E(u^*u^{*\prime}) = N^*\bar{\Lambda}$. y^* denotes the vector of numbers of accidents in which the different groups of motorists were involved in a year (cf. column 4 of Table 1): N^* is a diagonal matrix with on the main diagonal the total numbers of kilometres driven per year per group of motorists (equal to the products of the corresponding elements in columns 2 and 3 of Table 1); $\bar{X}$ is a matrix of order 15 x 2 (third-party insurance) or 8 x 2 (casco insurance) whose elements in the first column are all equal to 1, and whose elements in the second column equal the number 0,1, ..., 14 (third-party) or 0,1, ..., 7 (casco), which are the ages of the car distinguished; $\beta' = (\beta_1,\beta_2)$; and $\bar{\Lambda}$ is defined in (2.11).

Second, we assume that the number of accidents per (standard) month is linearly dependent on the age of the car. We use the regression model

$$(4.2)\quad y^* = M^*\bar{X}\alpha + v^*$$

with $E(v^*) = 0$ and $E(u^*u^{*\prime}) = M^*\bar{\Lambda}$.
y^* and $\bar{X}$ are defined as in the preceding paragraph; the diagonal elements of the diagonal matrix M^* are the numbers of months during which the motorists have driven, hence they are 12 times the elements of column 3 of Table 1; finally $\alpha' = (\alpha_1,\alpha_2)$, and $\bar{\Lambda}$ is the matrix as defined in (2.11).

Third, we analyse the relationship between the amounts of damage and the age of the car. Let the logarithm of the amount of damage depend linearly on the age of the car. After grouping of the observation we get the following regression model:

$$(4.3)\quad z^* = P^*\bar{X}\gamma + w^*$$

with u^* $N(0,D)$ distributed, where D has been defined in

(3.8), z^* denotes the vector of sums of the logarithms of the amounts of damage within each group; P^* is a diagonal matrix whose diagonal elements denote the numbers of accidents of the groups of motorists distinguished according to the age of the car (cf. column 4 of Table 1); the matrix $\bar{X}$ is defined above; and $\gamma' = (\gamma_1, \gamma_2)$.

The vectors β and α of models (4.1) and (4.2), respectively, are estimated by means of the iterative method described in Section 2. The iterative procedure for estimating β is stopped when

$$(4.4) \quad \frac{\hat{\beta}_i(r) - \hat{\beta}_2(r-1)}{\hat{\beta}_i(r)} \leq 0.0001 \quad \text{for } i = 1,2$$

where $\beta_i(r)$ stands for the estimate of β_i in the r^{th} cycle. A similar action is performed concerning the estimating of α.

In applying the least-squares estimation method, one generally computes the values of the correlation coefficient and the values of the standard errors of the point estimates of the parameters. In regard to the models presented in Section 2 these values have little meaning for testing purposes. Cramer (1964) shows that the correlation coefficient based on grouped observations will systematically exceed the one based on the individual observations. Koerts and Abrahamse (1970) studied problems arising from attempts to draw inferences by using the correlation coefficient in the general linear model. Moreover, our regression models do not include a constant term; this entails additional problems. For an analysis of the correlation coefficient in models with zero intercept we refer to Aigner (1971), Section 3.8.

If disturbances in a regression model are normally distributed, the values of the standard errors can be used in order to construct 95% confidence intervals for the parameters. In our models, in which the dependent variable is Poisson-distributed, the distribution of $\hat{\beta}$ cannot be derived easily. Moreover, if we use grouped

Table 2. Results of the regression to the number of accidents and the amount of damage of the age of the car (cf. equations (4.1), (4.2) and (4.3) successively).*

I model (4.1)	Average number of accidents per 18,000 km/year	$\hat{\beta}_1$ × 18,000	$\hat{\beta}_2$ × 18,000	Number of iterations	R^2
Third-party	0.1251	0.0865 (0.0061)	0.0108 (0.0015)	5	0.96
Casco	0.3209	0.2962 (0.0148)	0.0156 (0.0073)	4	0.99
II model (4.2)	Average number of accidents per motor car year	$\hat{\alpha}_1$	$\hat{\alpha}_2$	Number of iterations	R^2
Third-party	0.1127	0.1127 (0.0065)	0.0000 (0.0011)	∞	0.97
Casco	0.3530	0.3813 (0.0170)	-0.0158 (0.0067)	4	0.99
III model (4.3)	Average amount of damage in Dfl	$\hat{\gamma}_1$	$\hat{\gamma}_2$	s^2	R^2
Third-party	417	5.2954 (0.0831)	0.0017 (0.0130)	1.9217	0.99
Casco	459	5.2058 (0.0931)	0.0336 (0.0351)	2.8552	0.99

* The figures between brackets are the standard errors of the preceding point estimates of the parameters.

observations as in model (2.12), we must take into account that the variances of the estimates are larger after grouping (cf. Prais and Aitchison (1954)).

The results of the regression analysis are presented in Table 2. The squared correlation coefficient is then defined as

$$(4.5)\quad R^2 = 1 - \frac{\hat{u}^{*\prime}\hat{u}^*}{y^{*\prime}Ey^*}$$

with $\hat{u}^* = y^* - N^*\hat{X\beta}$ and $E = I - \frac{1}{n}\iota\iota'$; s^2 in model III is the estimate of the variance of the normally distributed disturbances. The number of cycles in the iteration procedure is generally small, namely 4 or 5 cycles, except for the third-party case of model II, for which we stopped the iterative procedure after 228 cycles. We believe that in this case the value of the explanatory variable has no effect on the value of the dependent variable, i.e. there is no relation between the number of accidents per month and the age of the car.

We observe that the impact of the age of the car on the casco damage is small, and negligible with respect to third-party damage. The standard errors corresponding to γ_2 are large, relative to the values of the point estimates. In spite of the high values of R^2 we conclude that the relation between the age of the car and the amount of damage is not significant.

Table 2 shows that the application of models (4.1) and (4.2) yield quite different results. For third-party insurance $\hat{\beta}_2$ in (4.1) is 0.0108, while $\hat{\alpha}_2$ in (4.2) is zero. The coefficients $\hat{\beta}_2$ and $\hat{\alpha}_2$ for casco are about equal in absolute value, but opposite in sign.

The coefficient $\hat{\alpha}_2$ in model (4.2) for casco is negative. This means that if the car is older than 24 years, the estimated numbers of casco accidents per month would become negative. This anomalous result is partly due to the fact that observations on casco-insured cars,

aged 8 years or more, are not available. For higher ages of the car a linear approximation of the relationship may no longer be justified. On the other hand, if a car insurance company wishes to use these estimates, in order to determine the premium, the result is not at all detrimental, however, since it will hardly ever happen that a motorist wishes to insure a 24-year-old car.

5. CONCLUDING REMARKS

In this paper we deal with problems connected with the estimation of the unknown parameters in a linear model in which the dependent variable

a) is Poisson-distributed,
b) assumes only a small number of different values and
c) may belong to different reference sets and in which
d) the number of observations is large.

We constructed a linear model which tackles these problems. The model has been successfully applied to observations on car accidents.

The analyses presented in this paper can be summarized as follows.

First. Basing the estimations on grouped observations has computational advantages. However, the price that has to be paid for these advantages is that decision-making becomes more difficult due to the loss of information.

Second. In the examples of Section 4 the iteration procedure required 4 or 5 cycles only; it need, however, not converge.

Third. The estimation procedure does not prevent the computed dependent variable from assuming negative

values. For each separate application one should consider whether this is a real problem. In many cases the probability that the explanatory variable exceeds some relatively high value is small enough to be disregarded. The method of grouping reduce the possibility of negative values of the computed dependent variable. Fourth. The method of grouping the observations, although derived for a single explanatory variable, can also be applied in the case of two or more explanatory variables. The criteria for grouping depend on the number of observations, the possibilities of classification per explanatory variable, and the number of explanatory variables.

REFERENCES

Aigner, D.J. (1971), Basic Econometrics, Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

C.B.S. (1965), "Het Bezit en Gebruik van Personenauto's in 1963", (English title: "Ownership and use of passenger cars in 1963"), Central Bureau of Statistics, The Hague.

C.B.S. (1967), "Het Bezit en Gebruik van Personenauto's in 1965", Central Bureau of Statistics, The Hague.

Chew, V. (1970), "Covariance Matrix Estimation in Linear Models", Journal of the American Statistical Association, 65, p. 173-181.

Cramer, J.S. (1964), "Efficient Grouping, Regression and Correlation in Engel Curve Analysis", Journal of the American Statistical Association, 59, p. 233-250.

Jorgenson, Dale W. (1961), "Multiple Regression Analysis of a Poisson Process", Journal of the American Statistical Association, 56, p. 235-245.

Koerts, J. and A.P.J. Abrahamse (1970), " The Correlation Coefficient in the General Linear Model", European Economic Review, 1, p. 401-427.

Prais, S.J. (1953), "A Note on Heteroscedastic Errors in Regression Analysis, with a Comment by H. Theil", Review of the International Statistical Institute, 21: 1/2, p. 28-29.

Prais, S.J. and J. Aitchison (1954), "The Grouping of Observations in Regression Analysis", Review of the International Statistical Institute, 22.

Rao, C.R. (1970), "Estimation of Heteroscedastic Variances in Linear Models", Journal of the American Statistical Association, 65, p. 161-172.

Van der Laan, B.S. (1971), A Note on the Probability Distributions of the Number and the Amount of the Claims of Motor Car Accidents, Report 7103 of the Econometric Institute, Netherlands School of Economics, Rotterdam.

Van der Laan, B S. and F.J.J. Boermans (1970), Some Probability Distributions in the Field of Motor Car Damages (An Empirical Study), Report 7001 of the Econometric Institute, Netherlands School of Economics, Rotterdam.

Weber, Donald C. (1971), "Accident Rate Potential: An Application of Multiple Regression Analysis of a Poisson Process", Journal of the American Statistical Association, 66, p. 285-288.