FACTORS AFFECTING FIRE LOSS—MULTIPLE REGRESSION MODELS WITH EXTREME VALUES

G. Ramachandran
U.K.

Summary

In his paper to the Tenth ASTIN Colloquium the author presented generalised extreme value techniques for making use of all large losses that are available for analysis and not merely the largest. In this paper the problem of assessing the relative contributions of various factors to fire losses is investigated. A model concerned with multiple regression with extreme observations of given rank is developed. It takes into consideration the biases due to the use of extremes and the differences between categories of risks in regard to the frequency of fires (or claims). By way of illustration the model was applied to the largest and second largest losses in the textile industries in the United Kingdom during the six-year period 1965 to 1970. The presence or absence of sprinklers, whether the buildings were single-storey or multi-storey, and total floor area were the independent variables included in this preliminary investigation. Judged from extreme losses sprinklers appear to reduce considerably the expected damage in all fires.

The technique enables different estimates to be obtained for each regression parameter for different ranks. It is desirable to have a single overall estimate for each parameter; and for this purpose a second model is developed for performing a regression analysis combining observations pertaining to a number of ranks. Covariances of the residual errors are also taken into account in this model.

Introduction

In my paper to the Tenth Colloquium [1] I presented generalised extreme value techniques for making use of all large losses that are available for analysis and not merely the largest. I dealt with methods of estimating the extreme value parameters and parameters

https://doi.org/10.1017/S051503610000934X Published online by Cambridge University Press

of an assumed form of parent distribution. In this paper the problem of assessing the relative contributions of various factors to fire losses is investigated. A multiple regression model with extreme observations of given rank is developed and illustrated with an example. A model for performing a combined regression analysis based on all large losses available is also described. As mentioned in the earlier paper [1] these methods would be particularly useful in situations where data are available only for large losses or claims.

The Standardised Variable

The losses or claims during a given period constitute a sample from a parent probability distribution. The variable z, i.e. loss, has a location parameter μ and scale parameter σ. The standardised variable

$$t = \frac{z - \mu}{\sigma} \tag{1}$$

has the (cumulative) distribution function $G(t)$ and density function $g(t)$. If the n losses in a period are arranged in decreasing order of magnitude the mth largest value of t from top is

$$t_{(m)n} = \frac{z_{(m)n} - \mu}{\sigma} \tag{2}$$

where $z_{(m)n}$ is the mth largest loss.

The probability density of $t_{(m)}$ is

$$\psi_m(y_m) = \frac{m^m}{(m - 1)\,!}\, e^{-my_m - me^{-y_m}} \tag{3}$$

if $G(t)$ is of the exponential type. The reduced variable

$$y_m = A_{mn}(t_{(m)n} - B_{mn}) \tag{4}$$

where A_{mn} and B_{mn} are solutions of

$$G(B_{mn}) = 1 - \frac{m}{n} \tag{5}$$

and

$$A_{mn} = \frac{n}{m}\, g(B_{mn}) \tag{6}$$

https://doi.org/10.1017/S051503610000934X Published online by Cambridge University Press

If the form of $G(t)$ is known, A_{mn} and B_{mn} could be calculated from (5) and (6). From (2) and (4),

$$\sigma = \sigma_{mz} \cdot A_{mn}/\sigma_m \tag{7}$$

and

$$\mu = \bar{z}_m - \sigma \left[B_{mn} + \frac{\bar{y}_m}{A_{mn}} \right] \tag{8}$$

where $\bar{z}_m$ and σ_{mz} are the expected value and standard deviation of $z_{(m)n}$ and $\bar{y}_m$ and σ_m the expected value and standard deviation of y_m.

Regression Based on *m*th Extreme

For a given set of values of p independent variables $v_i(i \doteq 1, \ldots p)$ the dependent variable z has an expected value μ_v and (residual) standard error σ_v. Also,

$$\mu_v = \beta_0 + \sum_{i=1}^{p} \beta_i v_i \tag{9}$$

The parameters μ and σ mentioned in the previous section take the values μ_v and σ_v in the regression model. The problem is to estimate the regression parameters $\beta_i(i = 0, 1, \ldots p)$ and σ_v using the mth largest observations $z_{(m)n}$ and the associated values of v_i.

Suppose the observations available for N periods are grouped into K independent categories depending upon the risk of fire loss, e.g. sprinklered, non-sprinklered and so on. Consider the model

$$z_{(m)jk} = \beta'_{0m} + \sum_{i=1}^{p} \beta_{im} v_{imjk} + e_{mjk} \tag{10}$$

where $z_{(m)jk}$ is the mth largest loss from top in the jth period $(j = 1, 2, \ldots N)$ for the kth category $(k = 1, 2, \ldots K)$ and v_{imjk} is the value of v_i associated with $z_{(m)jk}$. The residual error e_{mjk} has the expected value zero and its variance R^2_{mk} is known to be proportional to the variance σ^2_{mzk} of $z_{(m)jk}$. But from (7),

$$\sigma^2_{mzk} = \sigma^2_v \cdot \sigma^2_m / A^2_{mk} \tag{11}$$

where A_{mk} refers to the kth category with n_k number of fires or claims per period. The value of A_{mk} (and B_{mk}) could be obtained

https://doi.org/10.1017/S051503610000934X Published online by Cambridge University Press

from (5) and (6) with $n = n_k$. Since A_{mk} differs from category to category, the regression parameters β^1_{0m} and $\beta_{im} (i = 1, \ldots p)$ in (10) are to be estimated by minimising the weighted residual sum of squares

$$Q_m = \sum_{k=1}^{K} A^2_{mk} \sum_{j=1}^{N} \{z_{(m)jk} - \beta^1_{0m} - \sum_{i=1}^{p} \beta_{im}\, v_{imjk}\}^2 \qquad (12)$$

with respect to the parameters and solving the resulting normal equations.

For given values of the independent variables v_i the expected value of the mth largest loss would be given by

$$\mu^1_{vm} = \beta^1_{0m} + \sum_{i=1}^{p} \beta_{im}\, v_i \qquad (13)$$

If $\mu^1_{v\,mjk}$ is the value estimated by substituting in (13) the observed values v_{imjk} corresponding to the observed $z_{(m)jk}$, the weighted residual variance is

$$R^2_{mw} = \frac{1}{(NK - p - 1)} \sum_{k=l}^{K} A^2_{mk} \sum_{j=1}^{N} (z_{(m)jk} - \mu'_{vmjk})^2 \qquad (14)$$

Following the derivation of (7) it can be easily seen that

$$\sigma^2_{vm} = R^2_{mw} / \sigma^2_m \qquad (15)$$

The variance σ^2_{vm} is an estimate of σ^2_v for the parent regression defined in (9). From (8) the value of μ_v for the kth category is given by

$$\mu_{vmk} = \mu'_{vm} - \sigma_{vm} \left(B_{mk} + \frac{\bar{y}_m}{A_{mk}} \right)$$

which can be rewritten as

$$\mu_{vmk} = \beta_{0mk} + \sum_{c=1}^{p} \beta_{im} v_i \qquad (16)$$

where

$$\beta_{0mk} = \beta'_{0m} - \sigma_{vm} \left(B_{mk} + \frac{\bar{y}_m}{A_{mk}} \right) \qquad (17)$$

https://doi.org/10.1017/S051503610000934X Published online by Cambridge University Press

The values β_{im} ($i = 1, \ldots p$) are estimates of the parameters β_i of the parent regression shown in (9). The value β_{0mk} is an estimate of β_0 but varies from category to category. The values of $\bar{y}_m$ and σ^2_m have already been tabulated [2].

Application to Data

Data on large losses and the occupancies in which these fires occurred are available for a number of years. However, information on fire protection devices and other particulars of buildings involved in large fires is available only from 1965. For this reason the methods developed above were applied to the largest and second largest losses during the period 1965 to 1970. The textile industry in the United Kingdom was chosen as an illustration. A detailed account of this exercise was given in a recent Fire Research Note [3]. I shall now discuss the main features given in this note.

For an application of the asymptotic theory of extreme values the number of fires in a year in any category should be large and this requirement restricts the number of categories. Hence only four groups were considered; these were sprinklered and non-sprinklered in single-storey and multi-storey buildings. The top two losses in each of these groups during 1965 to 1970 were corrected for inflation, with 1965 as the base year. In the case of sprinklered buildings the figures referred to fires in which sprinklers operated. Of course, the probability of non-operation would be taken into account in a study of costs and benefits of sprinklers.

The presence or absence of sprinklers was denoted by the variable v_1. If the building was sprinklered, v_1 was assigned the value $+1$ and -1 if the building was not provided with sprinklers. Similarly, $+1$ was assigned to the variable v_2 if the building was multi-storeyed and -1 if it was single-storeyed. The interaction between the two factors, "sprinkler" and "storey", was not included in this study. Since the fire loss depends upon the size of the building [4, 5] the logarithm of the total floor area of the building was used as the third independent variable, v_3. The dependent variable was the logarithm of loss assuming the role of z in the previous sections. Previous studies [4, 5] indicate that the fire loss has a power relationship with the size of the building.

It was assumed that during the short period of six years there was

https://doi.org/10.1017/S051503610000934X Published online by Cambridge University Press

no appreciable increase in the number of fires and hence an average value was used for the sample size, n. About 50 per cent of the fires attended by fire brigades were small ones which did not spread beyond the article of origin [6]. Disregarding these fires which were of no economic importance the remaining 50 per cent in each of the four groups was used to denote the sample size for that group. It was assumed specifically that z has a normal distribution ie log normal for the actual loss. The values of A_{mk} and B_{mk} corresponding to sample sizes n_k were obtained from tables of the normal probability integral. The results are summarised in the following tables

TABLE 1

Results of regression analysis

Extremes (m)	$\beta^1{}_{0m}$	β_{1m}	β_{2m}	β_{3m}	$R^2{}_{mw}$	$\sigma^2{}_m$	$\sigma^2{}_{vm}$	$\bar{y}_m$
1	0.9813	–0.3262	0.0617	0.4262	0.7629	1.6449	0.4638	0.5772
2	1.5664	–0.3094	0.1972	0.1556	0.4987	0.6449	0.7733	0.2704

TABLE 2

Categories (k)	n	Extremes (m)	A_{mk}	B_{mk}	β_{0mk}	α_{mk}
Sprinklered Single-storey	125	1	2.7375	2.4089	–0.8027	–1.1906
		2	2.5000	2.1444	–0.4145	–0.9211
Sprinklered Multi-storey	250	1	2.9750	2.6521	–0.9569	–1.2214
		2	2.7375	2.4089	–0.6389	–0.7511
Non-sprinklered Single-storey	100	1	2.6700	2.3264	–0.7502	–0.4857
		2	2.4200	2.0538	–0.3380	–0.2258
Non-sprinklered Multi-storey	200	1	2.8800	2.5759	–0.9094	–0.5215
		2	2.6700	2.3264	–0.5685	–0.0619

https://doi.org/10.1017/S051503610000934X Published online by Cambridge University Press

Since the value of β_{0mk} varied, there were four regression equations for each extreme corresponding to the four groups. With v_1 and v_2 taking the values $+1$ or -1 the equation for the kth group was reduced to the following simple form with just v_3 as the independent variable.

$$\mu_{vmk} = \alpha_{mk} + \beta_{3m}v_3 \tag{18}$$

where

$$\alpha_{mk} = \beta_{0mk} + \beta_{1m}v_1 + \beta_{2m}v_2 \tag{19}$$

The values of α_{mk} are also given in Table 2.

For a log normal distribution the expected loss μ_x in the original units is

$$\mu_x = e^{\mu+\sigma^2/2}$$

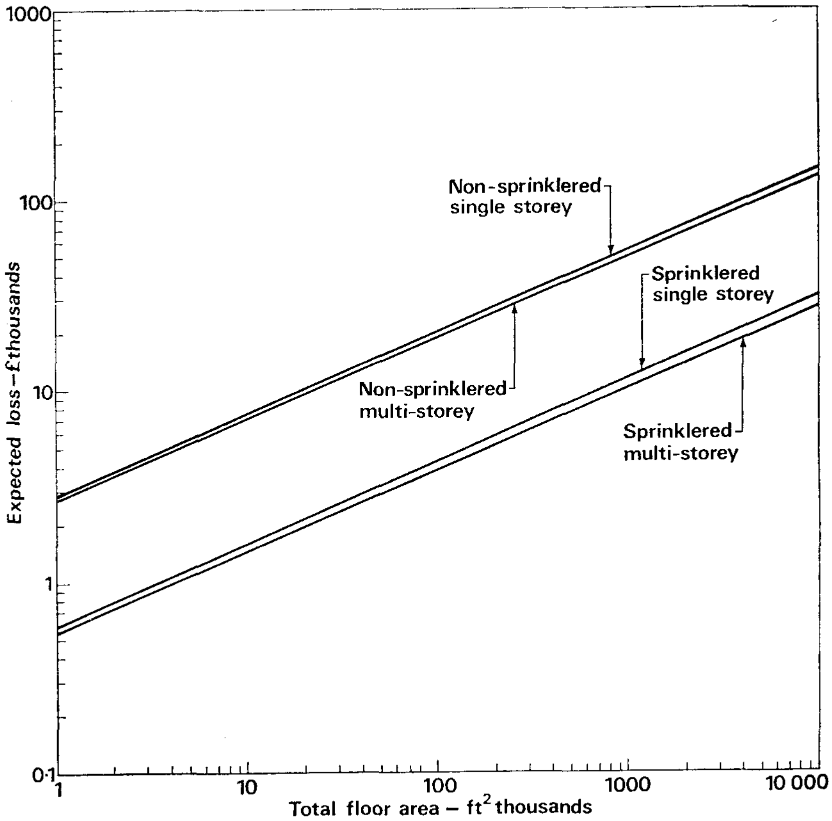

Fig. 1

https://doi.org/10.1017/S051503610000934X Published online by Cambridge University Press

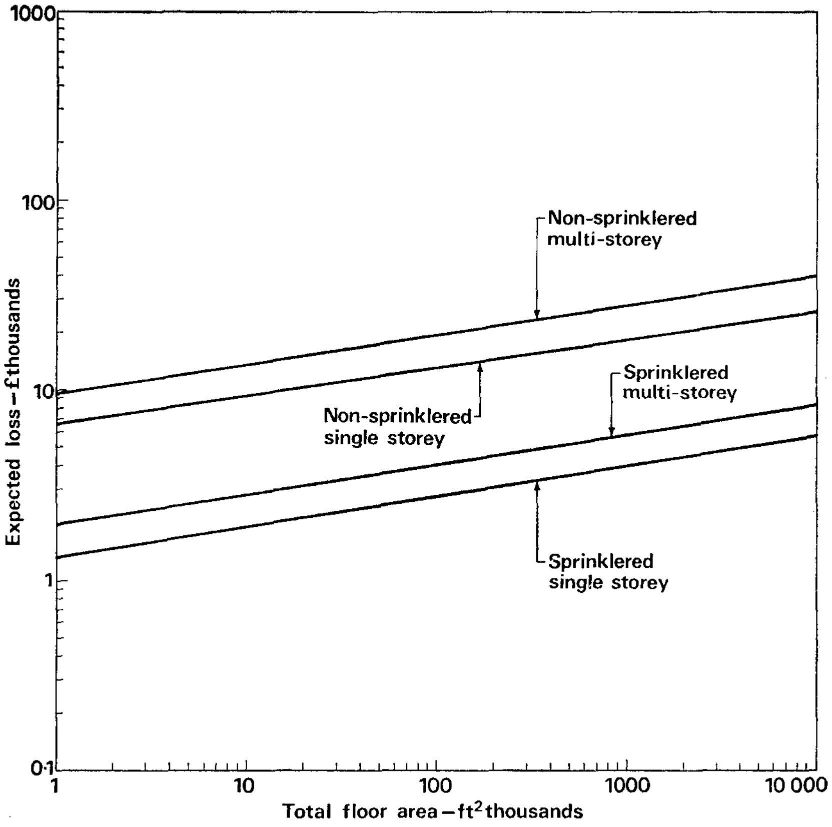

Fig. 2

where μ and σ are the mean and standard deviation of $z = \log_e x$ [7]. In the calculations 10 was used as base for z, the logarithm of loss in units of £'000. Hence the expected loss in the original units was

$$\mu_{xk} = 1000 \times e^{c\mu_{vmk} + c^2/2\ \sigma^2_{vm}} \qquad (20)$$

where $c = \log_e 10 = 2.3026$. Figures 1 and 2 depict the relationship (20) between the loss and total floor area for $m = 1$ and 2. The loss is at 1965 values.

For a given total floor area, the expected loss in a single-storey

https://doi.org/10.1017/S051503610000934X Published online by Cambridge University Press

building does not appear to differ very much from the expected loss in a multi-storey building. Perhaps, in a multi-storey building the horizontal spread of fire is restricted by better compartmentation but fire spreads vertically upwards. It is apparent that sprinklers reduce the expected loss to a considerable extent. From Fig. 1, for example, the expected loss in a fire in a building of total floor area 100 000 ft^2 would be about £20 000 if the building were not sprinklered but sprinklers would reduce the loss to £4000. The difference between the effect of sprinklers shown by Fig. 1 and Fig. 2 is due to random fluctuations.

Model for Combined Regression

For the kth population, from (2), (4), (10) and (17)

$$\bar{z}_{(m)k} = \beta_0 + \sum_{c=1}^{p} \beta_i \bar{v}_{imk} + \sigma_v \bar{t}_{(m)k} \tag{21}$$

where

$$\bar{z}_{(m)k} = \frac{1}{N} \sum_{j=1}^{N} z_{(m)jk}$$

$$\bar{v}_{imk} = \frac{1}{N} \sum_{j=1}^{N} v_{imjk}$$

and

$$\bar{t}_{(m)k} = B_{mk} + \frac{\bar{y}_m}{A_{mk}}$$

The variance of the residual error e_{mjk} is given by (11) while the covariance of e_{mjk} and e_{ljk} with $m > l$ is given by

$$\begin{aligned} &\sigma_v^2 \cdot cov\,[y_m, y_l]/A_{mk} \cdot A_{lk} \\ &= \sigma_v^2 \cdot \sigma_m^2/A_{mk} \cdot A_{lk} \end{aligned}$$

since the covariance of y_m and y_l with $m > l$ is the same as the variance of y_m[2]. Hence the error is σ_v^2 . V_k where the matrix V_k is of the form

https://doi.org/10.1017/S051503610000934X Published online by Cambridge University Press

$$\left|\begin{array}{ccccc}
\dfrac{\sigma_1^2}{A_{1k}^2} & \dfrac{\sigma_2^2}{A_{1k}\cdot A_{2k}} & \dfrac{\sigma_3^2}{A_{1k}\cdot A_{3k}} & \cdots\cdots & \dfrac{\sigma_r^2}{A_{1k}\cdot A_{rk}} \\
\dfrac{\sigma_2^2}{A_{2k}\cdot A_{1k}} & \dfrac{\sigma_2^2}{A_{2k}^2} & \dfrac{\sigma_3^2}{A_{2k}\cdot A_{3k}} & \cdots\cdots & \dfrac{\sigma_r^2}{A_{2k}\cdot A_{rk}} \\
\dfrac{\sigma_3^2}{A_{3k}\cdot A_{1k}} & \dfrac{\sigma_3^2}{A_{3k}\cdot A_{2k}} & \dfrac{\sigma_3^2}{A_{3k}^2} & \cdots\cdots & \dfrac{\sigma_r^2}{A_{3k}\cdot A_{rk}} \\
\cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\
\cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\
\dfrac{\sigma_r^2}{A_{rk}\cdot A_{1k}} & \dfrac{\sigma_r^2}{A_{rk}\cdot A_{2k}} & \dfrac{\sigma_r^2}{A_{rk}\cdot A_{3k}} & \cdots\cdots & \dfrac{\sigma_r^2}{A_{rk}^2}
\end{array}\right|$$

Then following Lloyd [8] we could obtain least squares estimates of $\beta_i (i = 0, 1, \ldots p)$ and σ_v by minimising the quadratic (matrix) form

$$(Z - C\Theta)'\, V^{-1}\, (Z - C\Theta) \qquad (22)$$

where

$$Z = \left|\begin{array}{c}
\bar{z}_{(1)1} \\ \bar{z}_{(2)1} \\ \cdots\cdots \\ \cdots\cdots \\ \bar{z}_{(r)1} \\ \bar{z}_{(1)2} \\ \bar{z}_{(2)2} \\ \cdots\cdots \\ \cdots\cdots \\ \bar{z}_{(r)2} \\ \cdots\cdots \\ \cdots\cdots \\ \bar{z}_{(1)K} \\ \bar{z}_{(2)K} \\ \cdots\cdots \\ \cdots\cdots \\ \bar{z}_{(r)K}
\end{array}\right| , \quad C = \left|\begin{array}{cccccc}
1 & \bar{v}_{111} & \bar{v}_{211} & \cdots\cdots & \bar{v}_{p11} & \bar{t}_{(1)1} \\
1 & \bar{v}_{121} & \bar{v}_{221} & \cdots\cdots & \bar{v}_{p21} & \bar{t}_{(2)1} \\
\cdots & \cdots & \cdots & \cdots\cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots\cdots & \cdots & \cdots \\
1 & \bar{v}_{1r1} & \bar{v}_{2r1} & \cdots\cdots & \bar{v}_{pr1} & \bar{t}_{(r)1} \\
1 & \bar{v}_{112} & \bar{v}_{212} & \cdots\cdots & \bar{v}_{p12} & \bar{t}_{(1)2} \\
1 & \bar{v}_{122} & \bar{v}_{222} & \cdots\cdots & \bar{v}_{p22} & \bar{t}_{(2)2} \\
\cdots & \cdots & \cdots & \cdots\cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots\cdots & \cdots & \cdots \\
1 & \bar{v}_{1r2} & \bar{v}_{2r2} & \cdots\cdots & \bar{v}_{pr2} & \bar{t}_{(r)2} \\
\cdots & \cdots & \cdots & \cdots\cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots\cdots & \cdots & \cdots \\
1 & \bar{v}_{11K} & \bar{v}_{21K} & \cdots\cdots & \bar{v}_{p1K} & \bar{t}_{(1)K} \\
1 & \bar{v}_{12K} & \bar{v}_{22K} & \cdots\cdots & \bar{v}_{p2K} & \bar{t}_{(2)K} \\
\cdots & \cdots & \cdots & \cdots\cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots\cdots & \cdots & \cdots \\
1 & \bar{v}_{1rK} & \bar{v}_{2rK} & \cdots\cdots & \bar{v}_{prk} & \bar{t}_{(r)K}
\end{array}\right|$$

https://doi.org/10.1017/S051503610000934X Published online by Cambridge University Press

$$\Theta = \begin{vmatrix} \beta_0 \\ \beta_1 \\ \\ \beta_p \\ \sigma_v \end{vmatrix} \quad \text{and } V = \begin{vmatrix} V_1 & 0 & 0 & 0..0 \\ 0 & V_2 & 0 & 0..0 \\ 0 & 0 & V_3 & 0..0 \\ & & & \ddots \\ 0 & 0 & 0 & 0 \cdot V_K \end{vmatrix}$$

It is assumed that observations are available for the top r extremes of K categories for N periods. The form (22) is a general linear model with correlated observations. It is not necessary to discuss here the estimation of parameters and other statistical problems. It is hoped to apply the combined regression model to actual data in the near future.

Discussion and Conclusions

Only large losses are available at present for multiple regression analysis to assess the effect of various factors on the expected damage. Hence the problem studied in this paper is to estimate the regression parameters by using extreme observations. Extreme values with any chosen rank, m, over successive periods could be used for this purpose. Estimates based on extremes would be biased since the entire range of the fire loss variable has not been covered. In the modified model presented in this paper, adjustments have been made to correct these biases.

In the example considered the main population was divided into independent categories. The largest and second largest losses in each category were used with years providing the replications. The number of fires per year in each category has to be large and this restricted the number of categories and parameters that could be included. It is possible to perform a separate regression analysis for each category but this would also restrict the number of parameters unless data over a large number of years were used. For these reasons a single regression analysis was carried out for each extreme, m.

It was assumed that the losses in the different categories for a given set of values of the regression variables v_i had independent but identical distributions viz. log normal. It was further assumed that

https://doi.org/10.1017/S051503610000934X Published online by Cambridge University Press

the distributions had different location parameters but a constant standard error σ_v. By including the parameters A_{mk} and B_{mk} the model takes into consideration the differences between categories in regard to the frequency of fires. The problem of confidence limits for the expected values and regression parameters is being investigated.

The estimates of regression parameters vary depending upon the rank, m, of extreme observations used. This variation is due to random fluctuations in the observations. It is also difficult to draw reliable conclusions from estimates based on just one extreme viz. the mth. For these reasons a combined regression model has been developed in this paper for using a number of extremes, say, $m = 1$ to r jointly and taking into consideration the variances and covariances of the residual errors.

Nelson and Hahn [9, 10] have discussed the linear estimation of a regression relationship from censored data using order statistics. In this paper similar estimation procedures are considered using extreme order statistics from large samples (asymptotic). It is possible to extend the model to extremes in small samples provided the moments of order statistics in such samples are either available or could be calculated. Teichroew [11] and Ruben [12] have dealt with order statistics from the normal distribution.

Acknowledgment

This paper is Crown Copyright, reproduced by permission of the Controller of Her Majesty's Stationery Office. It is contributed by permission of the Director, Building Research Establishment. (The Fire Research Station is the Joint Fire Research Organisation of the Department of the Environment and Fire Offices' Committee).

References

[1] Ramachandran, G. (1974), "Extreme value theory and large fire losses", *The ASTIN Bull.*, Vol. VII, Pt. 3. 293-310

[2] Ramachandran, G. (1972), "Extreme value theory and fire losses—further results", *Department of the Environment and Fire Offices' Committee Joint Fire Research Organisation Fire Research Note* No. 910.

[3] Ramachandran, G. (1973), "Factors affecting fire loss—Multiple regression model with extreme values", *Department of the Environment and Fire Offices' Committee Joint Fire Research Organisation Fire Research Note* No. 991.

https://doi.org/10.1017/S051503610000934X Published online by Cambridge University Press

[4] RAMACHANDRAN, G. (1970), "Fire loss indexes", *Ministry of Technology and Fire Offices' Committee Joint Fire Research Organisation Fire Researvh Note* No. 839.

[5] BLANDIN, A. (1956), "Bases techniques de l'assurance contre l'incendie", *A. Martel.*

[6] United Kingdom Fire Statistics. London. *Her Majesty's Stationery Office.* (Annual publication).

[7] BENKERT, L. G. (1963), "The log normal model for the distribution of one claim", *The ASTIN Bull.*, Vol. II, Pt. 1, 9-23.

[8] LLOYD, E. H. (1952), "Least squares estimation of location and scale parameters using order statistics", *Biometrika*, 39, 88-95.

[9] NELSON, W. and HAHN, G. J. (1972), "Linear estimation of a regression relationship from censored data. Part I. Simple methods and their application", *Technometrics*, 14, 247-269.

[10] NELSON, W. and HAHN, G. J. (1973), "Linear estimation of a regression relationship from censored data. Part II. Best linear unbiased estimation and theory", *Technometrics*, 15, 133-150.

[11] TEICHROEW, D. (1956), "Tables of expected values of order statistics and products of order statistics from samples of size 20 and less from the normal distribution", *Ann. Math. Statist.*, 27, 410-426.

[12] RUBEN, H. (1954), "On the moments of order statistics in samples from normal population", *Biometrika*, 41, 200-227.

https://doi.org/10.1017/S051503610000934X Published online by Cambridge University Press