



AgEcon SEARCH
RESEARCH IN AGRICULTURAL & APPLIED ECONOMICS

The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search
<http://ageconsearch.umn.edu>
aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

ECONOMETRIC INSTITUTE

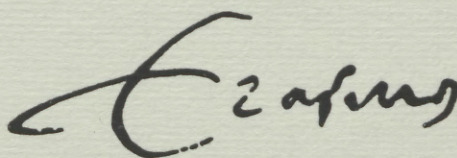
GIANNINI FOUNDATION OF
AGRICULTURAL ECONOMICS
LIBRARY
WITHDRAWN

JAN 14 1988

A STATISTICAL MODEL FOR THE EXPENSES
FOR MEDICAL SERVICES DURING A YEAR

B.S. VAN DER LAAN, J. KOERTS AND
J. REICHARDT

REPORT 8635/A



A STATISTICAL MODEL FOR THE EXPENSES FOR MEDICAL SERVICES DURING A YEAR

by

Bob van der Laan, Johan Koerts and Joke Reichardt

Abstract

The analysis of the total amount of expenses for medical services of an individual during a given year can be divided into two parts: the analysis of the probability of having non-zero expenses, and the analysis of the amount of the non-zero expenses. In this paper we construct a statistical model for the total amount of expenses for medical services. First, we analyse which factors influence the probability of having non-zero expenses on the basis of a logit model. Second, we analyse ten probability distributions that might be used to describe the amount of the non-zero expenses, where we distinguish five categories of non-zero expenses. Third, we examine which factors influence the parameters of the probability distribution of the amount of non-zero expenses of three categories on the basis of a lognormal model. The models are applied to Dutch health insurance data.

Contents

| | |
|--|----|
| 1. Introduction | 1 |
| 2. A general model for the total expenses for medical services | 2 |
| 3. A logit model for p | 4 |
| 4. Probability models for the amount of non-zero expenses for medical services | 6 |
| 5. Application | 11 |
| 5.1. The data | 11 |
| 5.2. Estimate of the parameters of the logit model | 15 |
| 5.3. Fitting probability distributions to the amount of non-zero expenses | 15 |
| 5.4. Estimate of the parameters of the models for the amount of non-zero expenses | 18 |
| 6. Summary and conclusions | 32 |
| References | 34 |
| Appendix | 36 |

1. Introduction^{1,2}

In this paper we analyze the costs for privately insured people of all expenses for hospital nursing, for clinical and for non-clinical services from specialists, and for certain additional costs during a given year. The aim is to derive a statistical model for the total expenses for medical services for an individual during a year. To test the model we have available insurance and claim data of 35,246 Dutch privately insured people, supplemented with additional personal data.

About 70 per cent of the Dutch population was medically insured with a type of national insurance (legally established by the Sick-Fund Insurance Act of 1966). People who are not insured via the national health insurance scheme, can insure themselves against the costs of medical care with a private insurance company. Insurances effected with an insurance company generally cover the costs of hospital nursing in one of the classes of hospital nursings³, the costs of clinical and non-clinical services of specialists, and some additional costs. An extensive description of this structure can be found in Van de Ven et al. (1980) and Van de Ven (1981).

Those aged under 16 years are always insured against the costs of hospital nursing in class III. Moreover, some information concerning personal data is irrelevant for young insurants. We therefore restrict our analysis to people above the age of 15 years. The number of people insured against the costs of hospital nursing in class I is relatively small, and will therefore not be analyzed. The general model however can also be applied to people aged under 16 years and to people with a class I insurance.

We found from earlier investigations that people insured with an insurance

-
1. The authors are indebted to the Stichting Het Zilveren Kruis (The Silver Cross Foundation) in Noordwijk, The Netherlands, which provided us with the insurance and claim data.
 2. An earlier version of this paper has been presented at the XIIIth International Biometric Conference, Seattle, Washington, U.S.A., July 27 - August 1, 1986.
 3. There are, in general, four classes of hospital nursing in Dutch hospitals, viz I, IIa, IIb and III, where the lowest, and least luxury one is class III.

against the costs of hospital nursing in class IIa or in class IIb can be considered to belong to one group of insured persons. In this paper we therefore only distinguish between people with a class II and people with a class III insurance.

It should be stressed that this is an analysis of the costs of health care for a select group. The most important characteristic of these insured persons is that their average annual family income is considerably higher than the average annual income of the total population in the Netherlands.

2. A general model for the total expenses on medical services

In theories concerning non-life insurances, the following assumptions are often made:

1. There is an unambiguous definition of the damage event.
2. The events are stochastically independent, i.e.
 - the occurrence of a certain event is independent of the occurrence of other events;
 - the amounts of damage from the different events are mutually independent;
 - the amount of damage of any certain event is independent of the number of events.

We use these assumptions in order to derive a probability distribution of the total costs of indemnity paid during a certain year to an individual. They are however unrealistic in the case of the costs of health care, since, for example,

- a. there may be interdependence between different events, i.e. cases of illness, and,
- b. from a statistical point of view, it is difficult to define what is a case of illness.

We use a different approach to analyze the costs of health care. We divide the analysis of the total costs of medical services of an individual in a given year into two parts:

- the analysis of the probability of having non-zero expenses: p , and
- the analysis of the amount of the non-zero expenses for medical services

during that year: Y .

Let $g(y; \theta)$ be the probability density function of Y , where θ is a vector of (unknown) parameters. Then the probability density function of the total amount of expenses for medical services of an individual during a given year, Z , equals

$$(2.1) \quad \begin{aligned} f(z; \theta) &= 1 - p && \text{if } z = 0 \\ &= p g(z; \theta) && \text{if } z > 0. \end{aligned}$$

Actually, $g(z; \theta)$ is the p.d.f. of a sum of random variables, each representing the medical expenses with respect to a different case of illness.

It may be statistically difficult to distinguish between cases of illness, but it is possible to classify the expenses for medical services into a number of different categories such as, for example, expenses for hospital nursing, for services of specialists or for maternity care. It can be expected that the probability p will be different for different categories of expenses for medical services. It can also be expected that the probability distribution function for the amount of non-zero expenses of one category differs from that of another category. Let there be c categories of medical services, then we get for each category k a probability p_k , and a probability density function f_k , where $k = 1, \dots, c$.

We shall assume for each k that the probability p depends on a vector of explanatory variables x , with corresponding vector of coefficients β . Hence

$$(2.4) \quad p_k = p_k(\beta) \quad k = 1, \dots, c.$$

In Section 3 we consider a logit model for each probability p_k .

The distribution of the random variables Y_k ($k = 1, \dots, c$) is unknown. In Section 4 we discuss some probability distributions for Y_k . Given the choice of the distribution function of Y_k , we assume for each k that the vector of parameters θ_k is dependent on a set of explanatory variables x , which reflects characteristics of an individual and of the type of his insurance, with

corresponding matrix of coefficients Ω , hence

$$(2.3) \quad \theta_k = h_k(\Omega) \quad k = 1, \dots, c.$$

3. A logit model for p

An extensive review of the developments in the area of qualitative response models has been given by, for example, Amemiya (1981) and McCullagh and Nelder (1983). For our problem the logit as well as the probit model seem plausible models. In this paper we choose for the logit model. This is not a serious restriction, because it has been shown (see Amemiya (1981)) that it is difficult to choose between probit and logit models, unless the number of observations is extremely large.

In our case the framework of the logit model can be formulated as follows. Assume that the probability, p , of having non-zero medical expenses depends on r explanatory variables x_1, \dots, x_r , where x_1 is taken to be unity. We assume p to be a function, F , of the linear combination $\sum \beta_j x_j$. Let $V_i = 1$ denote that the medical expenses of a given individual i are positive. Then we have

$$(3.1) \quad p_i = \Pr[V_i = 1] = F(\sum_j \beta_j x_{ij}).$$

Taking for F the logistic distribution function, we get

$$(3.2) \quad p_i = [1 + \exp(-\sum_j \beta_j x_{ij})]^{-1}.$$

To estimate β , we apply the method of maximum likelihood (abbreviated as ML). We therefore maximize the log likelihood function based on n individual observations

$$(3.3) \quad T_{ML} = \log\{L(\beta)\} = \sum_{i=1}^n [v_i \log\{F(\sum_j \beta_j x_{ij})\} + (1 - v_i) \log\{1 - F(\sum_j \beta_j x_{ij})\}]$$

with respect to β . The estimates of β_1, \dots, β_r can be determined with the help of a proper minimization routine⁴.

4. We used the NAG (Numerical Algorithms Group) minimization routine E04CCF, which is based on a simplex method.

In cases where the individuals are already classified in a limiting number of classes on the basis of a set of explanatory variables which have a categorical nature, we can simplify the likelihood function. Let the individuals be divided into K different classes, let n_k be the size of class k, and let m_k be the number of individuals in class k with non-zero expenses for medical services, then the log likelihood function can be written as

(3.4)

$$T_{ML} = \log\{L(\beta)\} = \sum_{k=1}^K [m_k \log\{F(\sum_j \beta_j x_{kj})\} + (n_k - m_k) \log\{1 - F(\sum_j \beta_j x_{kj})\}],$$

where the variables x_{kj} are dichotomous variables.

The values of the log likelihood function in the optimum can be used to choose between the different sets of explanatory variables. We prefer the set of explanatory variables which results in the highest value of T_{ML} , given the restriction that the contribution of each variable to the value of the likelihood function is significant with a significance level α . This set of variables will be called the optimal set of explanatory variables. To test whether the contribution of a variable to the value of the log likelihood function is significant, we apply the GLR (Generalized Likelihood Ratio) test. Note that minus the difference between the value of the log likelihood function at the null hypothesis $H_0: \beta_j = 0, \beta_i \neq 0, i = 1, \dots, r$ and $i \neq j$ and that at the alternative hypothesis $H_1: \beta_i \neq 0, i = 1, \dots, r$, is approximately chi-square distributed with 1 degree of freedom (cf., for example, Wilks (1962, p. 418)). This decision rule implies that the contribution of each variable of an optimal set of explanatory variables to the value of the log likelihood function is significant, and that, if we add an additional variable to an optimal set, the contribution of this variable to the log likelihood function is not significant, given a significance level α . It must be realized that variables, which are desirable on theoretical grounds, can be excluded by applying this decision rule.

A variable which has a categorical nature can be represented by one or more dichotomous variables. Variables which take only two different values can be represented by one dichotomous variable, and variables which take $d > 2$ different values can be represented by $d-1$ dichotomous variables. When one or more explanatory variables are represented by more than one dichotomous variable, the testing procedure must be slightly adapted. To test whether the

contribution of a variable is significant, we use the chi-square distribution with $d-1$ degrees of freedom. We shall take a value of α equal to 0.05 when we apply the ML-method in Subsection 5.2.

In general, the CPU time required by minimization routines increases with the degree of complexity of the likelihood function, with the number of parameters, and with the sample size, and can take high values. If we do not use the individual, but the classified observations, we gain much CPU time.

4. Probability models for the the amount of the expenses for medical services

There is no theoretical derivation of a distribution for the amount of the expenses for medical services. Beard et al. (1977, p. 18) state that "the existence of a distribution function $S(z)$ is an axiom in the theory of risk", where $S(z)$ stands for the distribution of the size of the claim for a certain loss. The choice of a distribution function will be determined by empirical considerations. Hogg and Klugman (1983) state that size-of-loss distributions in casualty insurance are often very long tailed skewed distributions. "Compounding explains why many of these losses have approximate Pareto, generalized Pareto, Burr and log-t distributions".⁵ They fit these probability distributions as well as the lognormal and the Weibull distribution to two data sets, one involving hurricane losses and one involving malpractice claims. Newhouse et al. (1980) suggest the Box-Cox family of distributions. They fit that family with success to a set of data involving medical expenses. The Box-Cox transformation is defined in such a way that, if the parameter λ of the transformation is equal to 0, we get a log transformation which results in a lognormal distribution. Moreover, it has been shown that when λ is equal to $1/3$, the resulting distribution is approximately a gamma distribution. The lognormal and the gamma distribution, therefore, are also considered as distributions for the amount of the medical expenses. Weber (1970) found that a compound exponential as well as a weighted sum of two exponential distributions provided a close fit to a sample of vehicle accident claims.

5. Hogg and Klugman (1983).

So we consider in Section 5.3 the following theoretical distributions:

- the exponential distribution,
- the gamma distribution,
- the Weibull distribution,
- a compound exponential distribution,
- the lognormal distribution,
- a weighted sum of two exponential distributions,
- the generalized Pareto distribution,
- the Burr distribution,
- the log-t distribution, and
- the normal distribution after a Box-Cox transformation.

The mathematical expressions of these distributions are given in the Appendix.

Even if we classify the data on expenses for medical services by categories, we have the problem that the data per category is not homogeneous. The parameters of the distributions will vary in principle from the one individual to another. To be able to examine for each category of expenses for medical services which distributions are the most proper ones as distributions for the amount of the non-zero expenses for medical services of an individual, we must construct groups of insured persons which are more or less homogeneous with respect to the particular category of medical services. Then we can fit probability distributions to the data of these groups.

As test criterium for the goodness of fit of any distribution, we consider the chi-square goodness-of-fit test statistic X^2 . If n is the total number of observations, k is the number of classes, n_j is the number of observations and e_j is the expected number of observations in class j under the null hypothesis, then X^2 is defined as

$$(4.1) \quad X^2 = \sum_{j=1}^k \frac{(n_j - e_j)^2}{e_j}.$$

If the parameters of the distributions to be tested are estimated with Best Asymptotic Normal estimators based on the likelihood function of the grouped data, X^2 is approximately chi-square distributed with $k-t-1$ degrees of freedom, where t is the number of parameters to be estimated. The p -value

based on this distribution gives a good indication of the degree of the goodness of the fit.

Then two classical problems must be solved. What must be the value of k , and how do we determine the class boundaries. The first problem will be solved by applying the formula of Mann and Wald (1942), modified by Williams (1950):

$$(4.2) \quad k \approx 2 \sqrt[5]{\frac{2(n-1)^2}{d^2}}, \quad \text{for } n \geq 42,$$

where d is given by the significance level α :

$$(4.3) \quad \alpha = \int_{-\infty}^d \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2} x^2\right\} dx.$$

We shall take as significance level $\alpha = 0.05$.

The other classical problem is to determine the class boundaries. Mann and Wald (1942), Gumbel (1942) and other authors suggest taking equiprobable classes, implying that the class boundaries are functions of the unknown parameters. Then X^2 can simply be written as

$$(4.4) \quad X_*^2 = \frac{k}{n} \sum_{j=1}^k n_j^2 - n.$$

We will apply this version of the chi-square test. We base the method on estimated equiprobable classes, for the class boundaries depend on the parameter estimates.

The value of X_*^2 depends on the method of estimating the parameters of the distribution. We will estimate the parameters by means of two different methods. When the original observations are available, the obvious method is to apply the method of maximum likelihood. Chernoff and Lehmann (1954) showed, however, that when applying the method of maximum likelihood, the limiting distribution of X^2 is not a chi-square distribution, but the distribution of a weighted sum of chi-square random variables, where the weights are in general

functions of the distribution to be tested. They also showed that the asymptotic distribution of X^2 lies in between a chi-square distribution with $k-t-1$ degrees of freedom and a chi-square distribution with $k-1$ degrees of freedom, where t is the number of parameters to be estimated. The value of X^2 , computed on the basis of maximum likelihood estimates will be denoted by X_{ML}^2 . Because we do not know the "exact" limiting distribution of X_{ML}^2 for all distributions we consider⁶, we compute the p-value on the basis of a chi-square distribution with $k-t-1$ degrees of freedom as well as on the basis of a chi-square distribution with $k-1$ degrees of freedom.

The other method of estimating is an adapted minimum chi-square method, which implies that we estimate the parameters on the basis of minimizing (4.4). Applying this method means that the test statistic X_{\star}^2 is a discrete random variable. Therefore, an infinity of values of the vector of unknown parameters θ will in general lead to a minimum value. From this set of values of the parameters we will choose that value of θ which maximizes the log likelihood function. The value of X_{\star}^2 , based on this estimation method will be denoted by X_{MC}^2 .

The exact distribution of X_{MC}^2 is unknown. Some preliminary investigations indicate that the sample size must be very large before the approximation of the exact distribution by a chi-square distribution is acceptable. A Monte Carlo study indicated that the p-value computed on the basis of a chi-square distribution with $k-t-1$ degrees of freedom is higher than the estimated p-value determined on the basis of the empirical distribution obtained by the Monte Carlo experiment. The p-values we present in Subsection 5.4 are computed on the basis of a $\chi^2(k-t-1)$ -distribution and must therefore be considered with care.

The maximum likelihood estimates cannot always be determined analytically. In these cases we apply a computer minimization routine. To estimate the parameters by means of this adapted minimum chi-square method, we always need

6. Dahiya and Gurland (1972) showed that in the case of continuous distributions with location and scale parameters, the null distribution of X_{ML}^2 does not depend on the parameters if these are estimated by sample mean and variance. In such a case the percentage points of the asymptotic null distribution of X_{ML}^2 can be determined, as has been done by Dahiya and Gurland for the normal distribution.

a minimization routine.⁷

Next, we assume that the distribution for the amount of a certain category for expenses for medical services of a subgroup of individuals which results in the best fit, is also the distribution of the amount of that category of expenses of the other individuals, except that the values of the parameters will be different. We assume that the parameters are functions of certain characteristics of the insurant and of his insurance.

It will appear that the lognormal distribution fits rather well to the amount of the expenses for medical services for the group of male insurants aged 40-64 years with a class III insurance, for three categories. For the sake of simplicity, as well as because of the clear interpretation of the parameters of this distribution, we assume that for three categories of expenses for medical services the amount of the expenses of a given individual i is lognormally distributed, with for each i the mean μ and the standard deviation σ functions of explanatory variables $\mu_i = \sum \omega_{1j} x_{ij}$ and $\sigma_i = \exp\{\sum \omega_{2j} x_{ij}\}$. The set of parameters $\Omega = \{\omega_{11}, \dots, \omega_{1r}, \omega_{21}, \dots, \omega_{2r}\}$ will be estimated by means of the method of maximum likelihood, hence we maximize the logarithm of

$$(4.5) \quad L(\omega_{11}, \dots, \omega_{1r}, \omega_{21}, \dots, \omega_{2r}) =$$

$$= \prod_{i=1}^m \frac{1}{y \exp\{\sum \omega_{2j} x_{ij}\} \sqrt{2\pi}} \exp\left\{-\frac{1}{2 \exp\{2 \sum \omega_{2j} x_{ij}\}} (\log y_i - \sum \omega_{1j} x_{ij})^2\right\}$$

with respect to Ω , where m denotes the number of insurants with non-zero expenses.

The decision rule to choose between different sets of explanatory variables and to test whether the contribution of each variable to the value of the log likelihood function is significant is analogous to that described in Section 3.

Finally, we estimate the values of the standard errors of the estimators of the parameters. These estimates are obtained from the Hessian matrix associated with the likelihood function. The values of the standard errors give an indication of the confidence intervals of the parameters.

7. Cf. footnote 4 on page 4.

5. Application

5.1. The data

The models presented in Sections 4 and 5 will be applied to insurance and claim data of 35,246 Dutch insurants with respect to the year 1976, which we obtained from the insurance company mentioned in footnote 2. These persons are insured privately against expenses for medical services, because they, or their head of family, have an annual income which is above a certain minimum amount. In 1976 the sphere of work of the insurance company was concentrated mainly in the western part of the country. Over 90 per cent of the insurants live in the four western provinces. We have available data about a number of independent variables concerning qualities of the insured persons and their insurance (see the detailed explanation of the variables below). Other variables, not in our data set, may also have an effect on the probability of having non-zero expenses or on the parameters of the probability distributions.

In 1976 the insurance company paid indemnification for all expenses for medical services for which one was insured: it was not possible to get a premium reduction on the basis of a deductible.

The procedure to fit the probability distributions has been carried out on the basis of the data of a subgroup of the 20,912 insured persons aged ≥ 16 years. For another subgroup of 14,334 insured persons aged ≥ 16 years we have additional personal data from an inquiry carried out among them. For the analysis which factors influence the parameters of the logit model for the probability of having non-zero expenses and the parameters of the lognormal distribution for the amount of non-zero expenses we used the data of the second group of insurants. Implicitly we assume that the results for the first group, as far as its probability distribution is concerned, can be generalized to the second group.

From the information available we consider the variables:

Expenses for medical services

Y_1 = the total amount of non-zero expenses for hospital nursing

Y_2 = the total amount of non-zero expenses for in-patient services from specialists

Y_3 = the total amount of non-zero expenses for out-patient services from specialists

Y_4 = the total amount of non-zero expenses for physiotherapy and non-zero expenses for expedients

Y_5 = the total amount of non-zero expenses for medical services abroad

Y_6 = the total amount of non-zero expenses for transportation services

Y_7 = the total amount of non-zero expenses for maternity care

Probability of having non-zero expenses for medical services

| | | |
|-----------|--------------|-------------------|
| $V_k = 1$ | if $Y_k > 0$ | |
| $= 0$ | if $Y_k = 0$ | $k = 1, \dots, 7$ |

Explanatory variables

insurant's age

| | |
|-----------------|-----------------|
| $x_2 = 1$ | ≥ 65 years |
| $x_3 = 1$ | 40-64 years |
| $x_2 = x_3 = 0$ | 16-39 years |

insurant's sex

| | |
|-----------|--------|
| $x_4 = 1$ | female |
| $= 0$ | male |

family

| | |
|-----------|---------------------|
| $x_5 = 1$ | member of a family |
| $= 0$ | a person on his own |

working hours

| | |
|-----------------|-----------|
| $x_6 = 1$ | part-time |
| $x_7 = 1$ | no job |
| $x_6 = x_7 = 0$ | full time |

type of job

| | |
|--------------------------|---|
| $x_8 = 1$ | no job |
| $x_9 = 1$ | wage-earner |
| $x_{10} = 1$ | self-employed |
| $x_8 = x_9 = x_{10} = 0$ | other, e.g. military service or co-operating member of the family |

place of residence

| | |
|-----------------------|--|
| $x_{11} = 1$ | the place of residence is an average sized town, or a large town |
| $x_{12} = 1$ | the place of residence is a dormitory town, or a small town |
| $x_{11} = x_{12} = 0$ | the place of residence is a rural district |

insured class of hospital nursing

| | |
|--------------|------------------|
| $x_{13} = 1$ | class IIA or IIB |
| $= 0$ | class III |

insurance against the cost of services of GP (GP insurance)

| | |
|--------------|-----|
| $x_{14} = 1$ | yes |
| $= 0$ | no |

The variables "working hours" and "type of job" are strongly associated with each other. If both variables appear to contribute significantly, we will not insert both variables separately as explanatory variables, but make the following combination:

working hours / type of job

$x_6 = 1$ and $x_7 = x_8 = x_9 = 0$
 $x_6 = 0$ and $x_7 = 1$
 $x_6 = 0$ and $x_7 = 0$
 $x_6 = 0$ and $x_8 = 1$
 $x_6 = 0$ and $x_9 = 1$
 $x_6 = 0$ and $x_8 = x_9 = 0$

no job
part-time
full time
wage earner
self-employed
other, e.g. military service or co-
operating member of the family

It is obvious that there is a one-to-one relationship between the probability of having non-zero expenses for hospital nursing and the probability of having non-zero expenses for in-patient services from specialists. We therefore take both categories of expenses together, so we consider the probability of having non-zero expenses for in-patient medical services. The other three categories are considered separately.

Various investigations, see, for example, Anderson (1968), Anderson et al. (1975), Manning et al. (1981) and Van de Ven and Van der Gaag (1982), analyse which factors influence the probability of having non-zero medical expenses and the amount of non-zero expenses. One of the most important factor which influence the probability of having non-zero expenses for medical services is the age of the individual. It can be stated that from a certain age onwards, the older the individual, the higher the probability. The sex of the individual is particularly important in the age group 16-39 years.

With respect to the influence of the family size on the probability of having non-zero medical services, it can be argued that the bigger the family size, the higher will be the probability.

Obviously, the number of hours one works per week and the type of job will be related in some way to the probability of having non-zero medical services. However, neither the causal direction of the relationship, nor the direction of the effect is clear. For example, one can be forced to accept a part-time job or to stop working because of a bad health. However, it can also be stated that some jobs are more dangerous than others. It is therefore possible that some full time jobs influence one's health negatively. We should realize this when considering the results.

A person's place of residence can influence the probability of having non-zero

medical services in two ways, for it can be stated that to live in some areas can be more healthy than to live in other areas. On the other hand, the extent of the health care facilities differs between the different areas.

We may expect that individuals who are frequently ill, and therefore frequently make use of medical services, are more inclined to effect an insurance against the costs of hospital nursing in class IIA or IIB. We may also expect that individuals who frequently make use of GP services are more inclined to effect a GP insurance. This remark must be kept in mind when one considers the results.

Obviously, the insurance class of hospital nursing will be a very important explanatory variable for the amount of the expenses for hospital nursing. We cannot say much about the effect of the other explanatory variables on the amounts of the expenses for medical services.

It appeared that there was too little data with respect to the categories "medical services abroad" and "maternity care". We therefore restrict our analysis to the other five categories.

Tables 1 and 2 give the number of insurants and the average amount of expenses for medical services by category of medical services and by explanatory variable for the second subgroup of 14,334 insured persons.

Table 1 shows that the number of insurants who claimed for expenses for hospital nursing is not equal to the number of insurants who claimed for expenses for in-patient services from specialists. This difference is due to the fact that fees for related services are not always paid during the same year. We will classify an insurant as a claimant if he has claimed either for expenses for hospital nursing or for expenses for in-patient services from specialists or for both.

The number of insurants with no job, classified according to "working hours", ought to be equal to the number of insurants with no job, classified according to "type of job". From Table 2 it appears that these number of insurants are unequal. This difference is caused by incorrect information supplied by the insurant inquired. When both variables are combined, we classify an insurant as having no job, when he gave "no job" either under the heading "working hours" or under the heading "type of job" or under both headings.

5.2. Estimate of the parameters of the logit models

We estimated the parameters of the logit model (3.1) for the four categories of medical services for several sets of explanatory variables, on the basis of the maximization of T_{ML} as given in (3.4), see Tables 3 and 4.

With respect to the category "in-patient medical services" and the category "out-patient services from specialists", we consider the groups of variables "working hours" and the group of variables "type of job" as one group of explanatory variables, represented by four dichotomous variables.

It appears that all variables, except the variables "type of job" and "place of residence", contribute significantly to the explanation of the probability of having non-zero expenses for in-patient medical services. From the p-values given in Table 4 it appears that the variable age is the most important explanatory variable.

All 7 groups of explanatory variables appear to contribute significantly to the explanation of the probability of having non-zero expenses for out-patient medical services from specialists. The values of the standard errors of the estimates of the coefficients are small in relation to the values of the estimates, except for the standard error of the estimate of variable x_9 (self-employed), which is approximately two third of the value of the estimate of the coefficient. From the p-values given in Table 4 we conclude that a relatively large contribution to the explanation of the probability is due to all variables except "family".

The model for physiotherapy and expedients contains three explanatory variables: "age", "insured class" and "GP insurance". The p-values indicate that all three variables make a relatively large contribution to the explanation of the probability.

The model for transportation services contains only two explanatory variables. If, however, we add the variable "place of residence" to the set of explanatory variables of this model, the p-value is only 0.0510. Hence, if we were to increase α a little, this variable would be contained in the optimal set. Again it appears that "age" is the most important variable.

5.3. Fitting probability distributions to the amount of non-zero expenses

In this subsection we analyze with regard to five categories of expenses for medical services, the results of the fit of ten probability distributions to

the data of the group of male insurants aged 40-64 years, who have a class III insurance.

Table 5 presents the frequency distributions of the amount of expenses. The table clearly shows that it is necessary to split the total amount of expenses into different categories. The amounts of expenses for hospital nursing in particular are much higher than the amounts of the expenses of the other categories.

The distributions of the first three categories are very long tailed skewed ones, the distributions of the last two have less thick tails. So it can be expected that for the first three categories the generalized Pareto, the Burr and the log-t distributions will show the best results, and that for the last two categories the gamma, the Weibull or the log normal distribution are more suitable.

There is a striking difference between the distribution of the amount of expenses for in-patient and that for out-patient services from specialists. The average amount of expenses for in-patient services from specialists is much higher than that for out-patient services from specialists. This difference can be partly explained by the fact that relatively many insurants visit a specialist only once or twice. One of the following amounts Dfl. 45.-, 48.-, 49.- or 50.- (about US \$ 20.-) occurred 130 times. Such an amount is the fee for one visit. None of the distributions considered were able to result in a good fit to these extremes. We therefore decided to fit conditional distributions to the amount of expenses for out-patient medical services, given that this amount is higher than Dfl. 100.-.

It is also noteworthy that there is a relatively high frequency of the amount of expenses for physiotherapy and expedients in the range Dfl. 500-599. This is caused by the fact that the amount of Dfl. 500.- occurred 25 times. None of the distributions could describe this particular feature of the data. It seems that this amount must correspond to a particular medical treatment. We could not trace whether this is in fact the case. Better fits were possible if we remove these 25 observations. The results given in Table 6 and 7 are based on this smaller set of observations.

Table 6 presents the parameter estimates and the values of the log likelihood function, and Table 7 gives the values of the test statistic and the p-values for each estimation method. Notice that the adapted minimum chi-square

estimation method results in considerable higher p-values than the maximum likelihood estimation method.

We shall now discuss the results of Table 7, using the p-values obtained by the ML-method.

(1) The exponential distribution does not result in reasonable fits for any category of expenses for medical services. The compound and the mixed exponential distribution result in a reasonable fit only for the category of out-patient services from specialists. These results suggest that the exponential distributions are not very suitable to describe the amount of expenses for medical services at least as far as our sample is concerned.

(2) The Weibull distribution is a boundary case: it describes two categories very well and one category reasonable well.

(3) The gamma and the lognormal distributions show good fits: the lognormal distribution describes three categories quite well, whereas the gamma distribution did well for four categories. The gamma distribution appears to be the best two-parameter distribution.

With respect to the three parameter distributions it appears that:

(4) the log-t distribution does not behave very well and the same holds true for the Box-Cox distribution;

(5) the generalized Pareto distribution behaves very well for four categories of expenses; the Burr distribution, however, shows the best fit of all distributions: the results are satisfactory for all categories of expenses.

If we consider the p-values obtained by applying the adapted minimum chi-square estimation method, the picture changes considerably. Now eight distribution result in reasonable to very good fits. More research with respect to this estimation method is needed. We therefore only consider the ML-method from now on.

In the next subsection we trace the factors which influence the parameters of the probability distribution for the amount of the non-zero expenses. Hence we must choose a proper distribution for the amount of the expenses. In order to perform such an analysis it is desirable that the parameters have a clear interpretation and that the estimation procedure can be carried out simply. As a first step we take the lognormal distribution for this purpose, because it is the most suitable one to carry out such an analysis. Furthermore, we limit

our analysis to the first three categories of medical services. Finally, we assume for each of the three categories that the parameters μ and σ are functions of the explanatory variables which represent the characteristics of the insured person and his insurance.

5.4. Estimate of the parameters of the models for the amount of non-zero expenses

We estimated the parameters of the likelihood function as given in (3.5) for the three categories of medical services for several sets of explanatory variables. The results are given in Tables 8 and 9. In Table 8 we present the values of the estimates of the coefficients with their standard errors concerning the optimal sets of explanatory variables, for μ as well as for σ . In Table 9 we give the p-values for the sets of explanatory variables which are constructed from the optimal sets by omitting one variable from or by adding one variable to these sets.

With respect to the category of expenses for hospital nursing it appears that the variables "age", "sex" and "insured class" in the relationship for μ and the variables "age", "working hours" and "place of residence" in the relationship for σ contribute significantly to the value of the log likelihood function. However, the values of the standard errors of the estimates of the coefficients of the variables x_3 (age 40-64 years) and x_7 (no job) in the relationship for σ suggest that these estimates do not differ significantly from zero.

Concerning the category of expenses for in-patient services from specialists we observe that the variables "age", "place of residence" and "insured class" in relationship for μ and the variables "family" and "insured class" in the relationship for σ contribute significantly to the value of the log likelihood function. We remark that this model does not contain a constant term in the relationship for σ . The values of the standard errors of the estimates of the coefficients of the variables x_{11} and x_{12} (place of residence) in the relationship for μ suggest that these estimates do not differ significantly from zero.

It could be expected that the coefficient of the variable "insured class" in the relationship for μ is positive for these two categories of expenses. For, the fees for hospital nursing and for in-patient services from specialists were in 1976 higher for insurants with a class II insurance than those for

insurants with a class III insurance. The insured class does not influence the fees for services for the other category of expenses. We therefore do not expect a significant contribution of the insured class to the value of the log likelihood function concerning the relationship for μ .

The variables "age" and "working hours / type of job" in the relationship for μ , and the variable "insured class" in the relationship for σ appear to contribute significantly to the value of the log likelihood function for expenses for out-patient services from specialists. The estimates of all coefficients differ significantly from zero, given a significance level of 0.05.

In Table 10 we present the values of the log likelihood function and the p-values in the case that the parameters μ and σ are not functions of the explanatory variables. These p-values are very small, implying that the optimal sets of explanatory variables result in a very large contribution to the value of the log likelihood function.

Table 1. Number of insurants and average amounts of expenses for medical services by category of medical services.

| Category | Number of claimants (*) | | Average amount of non-zero medical expenses in Dfl. |
|--|-------------------------|----------|---|
| | absolute | relative | |
| Y ₁ : hospital nursing | 1254 | 0.0875 | 4476 |
| Y ₂ : in-patient services from specialists | 1204 | 0.0840 | 1096 |
| Y ₃ : out-patient services from specialists | 6217 | 0.4337 | 322 |
| Y ₄ : physiotherapy and expedients | 834 | 0.0582 | 367 |
| Y ₅ : medical services abroad | 63 | 0.0044 | 186 |
| Y ₆ : transportation services | 396 | 0.0276 | 274 |
| Y ₇ : maternity care | 222 | 0.0155 | 1384 |
| all categories | 6648 | 0.4638 | 1454 |

*) The total number of insurants equals 14334.

Table 2. Number of insurants and average amounts of expenses for medical services by explanatory variable.

| Variable | Number of insurants | Number of claimants | | Average amount of non-zero medical services in Dfl. |
|----------------------------------|---------------------|---------------------|----------|---|
| | | absolute | relative | |
| insurant's age | | | | |
| 16-39 years | 6251 | 2431 | 0.3889 | 1052 |
| 40-64 years | 6091 | 3064 | 0.5030 | 1436 |
| < 65 years | 1992 | 1153 | 0.5788 | 2351 |
| insurant's sex | | | | |
| female | 7033 | 3554 | 0.5053 | 1533 |
| male | 7301 | 3094 | 0.4238 | 1363 |
| family | | | | |
| member | 12495 | 5772 | 0.4619 | 1414 |
| a person on his own | 1839 | 876 | 0.4763 | 1720 |
| working hours | | | | |
| full time | 5869 | 2445 | 0.4166 | 1161 |
| part-time | 1040 | 508 | 0.4885 | 1509 |
| no job | 7425 | 3695 | 0.4976 | 1641 |
| type of job | | | | |
| wage earner | 4534 | 1985 | 0.4378 | 1128 |
| self-employed | 1919 | 768 | 0.4002 | 1350 |
| other | 627 | 249 | 0.3971 | 1547 |
| no job | 7254 | 3646 | 0.5026 | 1647 |
| place of residence | | | | |
| rural district | 1813 | 1568 | 0.4112 | 1559 |
| dormitory town or small town | 5638 | 2583 | 0.4581 | 1492 |
| average sized town or large town | 4883 | 2497 | 0.5114 | 1349 |
| insured class | | | | |
| class III | 10638 | 4591 | 0.4316 | 1224 |
| class IIA or IIB | 3696 | 2057 | 0.5565 | 1968 |
| GP insurance | | | | |
| yes | 8677 | 3781 | 0.4357 | 1412 |
| no | 5657 | 2867 | 0.5068 | 1510 |
| total | 14334 | 6648 | 0.4638 | 1454 |

Table 3. Parameter estimates of the logit models, with standard errors between brackets.

| Explanatory variable | In-patient medical services | Out-patient services from specialists | Physiotherapy and expedients | Transportation services |
|--|-----------------------------|---------------------------------------|------------------------------|-------------------------|
| constant term | | | | |
| $x_1 = 1$ | -3.1163 (0.1091) | -1.5886 (0.1303) | -3.5687 (0.0798) | -4.1686 (0.1092) |
| age | | | | |
| $x_2 = 1 (\geq 65)$ | 0.6290 (0.0917) | 0.6948 (0.0600) | 0.7634 (0.1135) | 1.3431 (0.1336) |
| $x_3 = 1 (40-64)$ | 0.3600 (0.0665) | 0.4992 (0.0390) | 0.7759 (0.0885) | 0.4547 (0.1260) |
| sex | | | | |
| $x_4 = 1$ (female) | 0.3050 (0.0713) | 0.2336 (0.0434) | | |
| family | | | | |
| $x_5 = 1$ (member) | 0.2133 (0.0917) | 0.1343 (0.0557) | | |
| working hours | | | | |
| $x_6 = 1$ (part-time) | 0.3132 (0.1182) | | | |
| $x_7 = 1$ (no job) | 0.1829 (0.0819) | | | |
| working hours / type of job | | | | |
| $x_6 = 1$ (no job) | | 0.4901 (0.1132) | | |
| $x_7 = 1$ (full time) | | 0.2568 (0.0813) | | |
| $x_8 = 1$ (wage earners) | | 0.3722 (0.1122) | | |
| $x_9 = 1$ (self-employed) | | 0.1851 (0.1184) | | |
| place of residence | | | | |
| $x_{11} = 1$ (average sized or large town) | | 0.3663 (0.0454) | | |
| $x_{12} = 1$ (dormitory town or small town) | | 0.1678 (0.0439) | | |
| insured class | | | | |
| $x_{13} = 1$ (class II) | 0.2532 (0.0644) | 0.2572 (0.0417) | 0.4237 (0.0781) | |
| GP insurance | | | | |
| $x_{14} = 1$ (yes) | 0.2019 (0.0579) | 0.2388 (0.0360) | 0.3344 (0.0725) | 0.2825 (0.1031) |
| Average | 0.0981 | 0.4337 | 0.0582 | 0.0276 |
| Variance | 0.0885 | 0.2456 | 0.0548 | 0.0269 |
| Log likelihood | -4505.60 | -9486.43 | -3089.78 | -1757.89 |
| Number of observations | 14334 | 14334 | 14334 | 14334 |

Table 4. P-values with respect to the logit models, for several sets of explanatory variables.

| Omitted / added explanatory variable (*) | In-patient medical services | Out-patient services from specialists | Physiother- apy and expedients | Transpor- tation services |
|---|-----------------------------------|---|--------------------------------------|---------------------------------|
| <u>Omitted</u> | | | | |
| age: x_2, x_3 | 0.1045×10^{-11} | 0.0000 | 0.8468×10^{-19} | 0.4716×10^{-21} |
| sex: x_4 | 0.1653×10^{-4} | 0.7107×10^{-7} | | |
| family: x_5 | 0.1818×10^{-1} | 0.1581×10^{-1} | | |
| type of job: x_6, x_7 | 0.1829×10^{-1} | | | |
| working hours / type of job: x_6, x_7, x_8, x_9 | | 0.6389×10^{-6} | | |
| place of residence: x_{11}, x_{12} | | 0.2988×10^{-14} | | |
| insured class: x_{13} | 0.9737×10^{-4} | 0.6926×10^{-9} | 0.8094×10^{-7} | |
| GP insurance: x_{14} | 0.5212×10^{-3} | 0.3205×10^{-10} | 0.4376×10^{-5} | 0.6350×10^{-2} |
| <u>Added</u> | | | | |
| sex: x_4 | | | 0.3067 | 0.9670 |
| family: x_5 | | | 0.3743 | 0.6902 |
| working hours: x_6, x_7 | | | 0.1488 | 0.0995 |
| type of job: x_8, x_9, x_{10} | 0.1274 | | 0.2008 | 0.3053 |
| place of residence: x_{11}, x_{12} | 0.8661 | | 0.2792 | 0.0510 |
| insured class: x_{13} | | | | 0.1766 |

*) First, each of the explanatory variables, or group of explanatory variables, from the optimal set of explanatory variables, given in Table 3 is (are) omitted from this set one by one. Second, each of the explanatory variables, or group of explanatory variables, which do(es) not belong to the optimal set is (are) added to this set.

Table 5. Frequency distributions of the amounts of non-zero expenses for medical services for the group of male insurants, aged 40-64 years, with class III insurance, for five categories of expenses for medical services.

| Range in Dfl. | Hospital nursing | Range in Dfl. | In-patient services from specialists | Out-patient services from specialists | Range in Dfl. | Physiotherapy and expedients | Transportation services |
|--------------------|------------------|---------------|--------------------------------------|---------------------------------------|---------------|------------------------------|-------------------------|
| 0 - 999 | 44 | 0 - 99 | 6 | 418 | 0 - 99 | 34 | 22 |
| 1000 - 1999 | 47 | 100 - 199 | 16 | 254 | 100 - 199 | 42 | 31 |
| 2000 - 2999 | 39 | 200 - 299 | 18 | 153 | 200 - 299 | 27 | 12 |
| 3000 - 3999 | 32 | 300 - 399 | 20 | 117 | 300 - 399 | 32 | 7 |
| 4000 - 4999 | 19 | 400 - 499 | 20 | 86 | 400 - 499 | 19 | 5 |
| 5000 - 5999 | 9 | 500 - 599 | 12 | 51 | 500 - 599 | 39 | 4 |
| 6000 - 6999 | 7 | 600 - 699 | 23 | 46 | 600 - 699 | 7 | 1 |
| 7000 - 7999 | 12 | 700 - 799 | 10 | 30 | 700 - 799 | 4 | 1 |
| 8000 - 8999 | 3 | 800 - 899 | 9 | 28 | 800 - 899 | 0 | 0 |
| 9000 - 9999 | 6 | 900 - 999 | 10 | 18 | 900 - 999 | 4 | 0 |
| 10000 - 10999 | 5 | 1000 - 1099 | 8 | 14 | ≥ 1000 | 5 | 2 |
| 11000 - 11999 | 4 | 1100 - 1199 | 7 | 16 | | | |
| 12000 - 12999 | 3 | 1200 - 1299 | 10 | 9 | | | |
| 13000 - 13999 | 1 | 1300 - 1399 | 9 | 6 | | | |
| 14000 - 14999 | 2 | 1400 - 1499 | 6 | 6 | | | |
| 15000 - 19999 | 4 | 1500 - 1599 | 7 | 6 | | | |
| 20000 - 24999 | 4 | 1600 - 1699 | 4 | 3 | | | |
| 25000 - 29999 | 4 | 1700 - 1799 | 2 | 4 | | | |
| ≥ 30000 | 5 | 1800 - 1899 | 6 | 2 | | | |
| | | 1900 - 1999 | 3 | 0 | | | |
| | | 2000 - 2499 | 9 | 4 | | | |
| | | 2500 - 2999 | 6 | 4 | | | |
| | | ≥ 3000 | 13 | 3 | | | |
| Total | 250 | | 234 | 1278 | | 213 | 85 |
| Average | 5200.72 | | 1057.16 | 325.05 | | 369.56 | 223.34 |
| Standard deviation | 7580.84 | | 1058.16 | 414.59 | | 422.22 | 212.42 |

Table 6. Parameter estimates and values of the log likelihood for different distribution functions, for the group of insured persons aged 40-64 years with class III insurance, for five categories of expenses for medical services.

| Distribution | (1) | (2) | Hospital nursing | In-patient services from specialists | Out-patient services from specialists | Physiother- apy and expedients | Transpor- tation services |
|-------------------------|-----|------------|--------------------------|---|--|--------------------------------------|---------------------------------|
| Exponential | ML | β | 5230.72 | 1057.16 | 558.12 | 352.22 | 223.34 |
| | | LL | -2390.58 | -1863.42 | -6284.48 | -1290.48 | -544.74 |
| | MC | β | 4504.28 | 969.68 | 334.35 | 380.51 | 276.34 |
| | | LL | -2393.52 | -1864.32 | -6419.10 | -1291.03 | -546.54 |
| Gamma | ML | α | 0.8608 | 1.3639 | 0.3980 | 1.3112 | 1.3179 |
| | | β | 6076.39 | 775.08 | 569.10 | 268.63 | 169.47 |
| | | LL | -2388.64 | -1857.07 | -5888.22 | -1286.53 | -542.89 |
| | MC | α | 1.0810 | 1.4109 | 0.3729 | 1.5565 | 1.2823 |
| | | β | 4058.61 | 711.58 | 580.68 | 198.25 | 176.54 |
| | | LL | -2397.65 | 1857.60 | -5888.26 | -1290.99 | -542.92 |
| Weibull | ML | τ | 0.8682 | 1.1422 | 0.7446 | 1.0750 | 1.1440 |
| | | β | 1575.10 | 3021.02 | 59.1555 | 566.31 | 515.75 |
| | | LL | -2385.37 | -1859.85 | -5886.55 | -1289.49 | -543.46 |
| | MC | τ | 0.9927 | 1.1352 | 0.7662 | 1.1703 | 1.1544 |
| | | β | 4288.14 | 2734.69 | 71.8244 | 834.79 | 551.62 |
| | | LL | -2392.30 | -1860.13 | -5887.33 | -1296.09 | -543.47 |
| Compound Exponential | ML | α | 3.5541 | 1795.56 | 5.1006 | 12.0907 | 3508.4×10^6 |
| | | β | 13294.1 | 1897.1×10^3 | 1370.43 | 3881.48 | 7835.6×10^8 |
| | | LL | -2377.08 | -1863.42 | -5885.88 | -1288.60 | -542.64 |
| | MC | α | 4.8336 | 2110.02 | 5.7480 | 13.5365 | 3789.1×10^6 |
| | | β | 16653.3 | 2045.5×10^3 | 1660.07 | 3636.52 | 7810.4×10^8 |
| | | LL | -2379.48 | -1864.32 | -5886.57 | -1291.59 | -542.92 |
| Lognormal | ML | μ | 7.8794 | 6.5540 | 5.5322 | 5.4367 | 4.9836 |
| | | σ | 1.2727 | 0.9651 | 0.9537 | 0.9617 | 1.0212 |
| | | LL | -2384.86 | -1857.35 | -5887.79 | -1281.57 | -546.00 |
| | MC | μ | 7.8864 | 6.4999 | 5.4680 | 5.4660 | 5.2052 |
| | | σ | 1.0807 | 0.9531 | 0.9738 | 0.8893 | 0.7686 |
| | | LL | -2392.35 | -1857.76 | -5888.39 | -1282.89 | -557.90 |
| Mixed Exponential | ML | α | 0.2477×10^{-10} | 0.9990 | 0.1994 | 0.0262 | 0.1494×10^{-10} |
| | | θ_1 | 741.57 | 1057.20 | 679.95 | 1973.82 | 1019.07 |
| | | θ_2 | 5230.73 | 1026.47 | 255.85 | 308.65 | 223.04 |
| | | LL | -2390.58 | -1863.42 | -5886.90 | -1284.71 | -544.74 |
| | MC | α | 0.0784 | 0.9988 | 0.1935 | 0.0110 | 0.1665×10^{-12} |
| | | θ_1 | 4678.98 | 1141.01 | 669.95 | 2032.94 | 2052.67 |
| | | θ_2 | 4488.55 | 1039.86 | 232.87 | 317.50 | 276.34 |
| | | LL | -2393.35 | -1864.09 | 5888.57 | -1285.20 | -546.54 |

Table 6. Continued.

| Distribution | (1) | (2) | Hospital nursing | In-patient services from specialists | Out-patient services from specialists | Physiother- apy and expedients | Transpor- tation services |
|-----------------------|-----|-----------|---------------------|---|--|--------------------------------------|---------------------------------|
| Generalised Pareto | ML | λ | 6333.92 | 2441.16 | 1999.16 | 689.10 | 978.93 |
| | | k | 1.3514 | 1.8373 | 0.8066 | 1.9222 | 1.5588 |
| | | α | 2.6233 | 5.2424 | 6.1936 | 4.8397 | 7.8313 |
| | | LL | -2374.06 | -1852.41 | -5885.65 | -1276.74 | -542.23 |
| | MC | λ | 6211.67 | 2714.26 | 1964.17 | 687.02 | 1138.20 |
| | | k | 1.4732 | 1.6319 | 0.8078 | 1.9482 | 1.4827 |
| | | α | 2.6529 | 4.8966 | 5.9659 | 4.8726 | 8.0458 |
| | | LL | -2374.91 | -1853.48 | -5885.87 | -1276.75 | -542.48 |
| Burr | ML | λ | 52129.88 | 65902.80 | 1216.98 | 16436.89 | 5075.27 |
| | | τ | 1.2889 | 1.5244 | 0.9095 | 1.5924 | 1.4216 |
| | | α | 1.5930 | 2.1798 | 7.7497 | 2.0607 | 2.8619 |
| | | LL | -2373.31 | -1852.64 | -5885.68 | -1275.29 | -542.07 |
| | MC | λ | 50425.54 | 55107.94 | 1293.31 | 18097.20 | 4771.32 |
| | | τ | 1.2554 | 1.5043 | 0.8980 | 1.5781 | 1.3413 |
| | | α | 1.7571 | 2.0416 | 8.4335 | 2.4411 | 3.6311 |
| | | LL | -2374.04 | -1852.91 | -5886.24 | -1275.61 | -542.33 |
| Log-Student | ML | μ | 7.9434 | 6.5849 | 5.5321 | 5.4951 | 5.0739 |
| | | τ | 0.9680 | 0.8688 | 0.9537 | 0.8107 | 0.7895 |
| | | v | 4.5617 | 10.5832 | 0.7139 $\times 10^7$ | 6.6182 | 4.4854 |
| | | LL | -2373.78 | -1855.33 | -5887.79 | -1278.78 | -543.96 |
| | MC | μ | 7.8852 | 6.6239 | 5.4674 | 5.4654 | 5.1865 |
| | | τ | 1.0174 | 0.9298 | 0.9770 | 0.7948 | 0.6740 |
| | | v | 5.0836 | 11.3817 | 3021.37 | 8.3007 | 3.2982 |
| | | LL | -2374.27 | -1856.25 | -5888.37 | -1279.25 | -545.05 |
| Box-Cox | ML | λ | 0.1354 | 0.1575 | 0.1746 | 0.1309 | 0.2233 |
| | | μ | 14.3876 | 11.6777 | 8.7096 | 8.0457 | 9.4878 |
| | | σ | 3.5753 | 2.6563 | 3.0783 | 1.9292 | 2.9762 |
| | | LL | -2376.34 | -1852.70 | -5885.83 | -1278.64 | -542.33 |
| | MC | λ | 0.1377 | 0.1515 | 0.1745 | 0.1483 | 0.2108 |
| | | μ | 14.6992 | 11.5047 | 8.7242 | 8.5262 | 9.2414 |
| | | σ | 3.3303 | 2.6500 | 2.9122 | 1.8675 | 3.0497 |
| | | LL | -2378.70 | -1853.20 | -5887.74 | -1282.04 | -543.01 |

(1) Estimation method.

(2) Parameters and log likelihood function.

Table 7. Values of the test statistic and p-values (with between brackets the number of d.f.) for different distribution functions, for the group of male insured persons aged 40-64 years with class III insurance, for five categories of expenses for medical services.

| Distribution | (1) | Hospital nursing | In-patient services from specialists | Out-patient services from specialists | Physiotherapy and expedients | Transportation services |
|----------------------|-----|------------------|--------------------------------------|---------------------------------------|------------------------------|-------------------------|
| Exponential | 1 | 33.97 | 35.09 | 202.67 | 31.73 | 20.99 |
| | 2 | 0.0055 (16) | 0.0039 (16) | 0.0000 (27) | 0.0044 (14) | 0.0212 (10) |
| | 3 | 0.0034 (15) | 0.0024 (15) | 0.0000 (26) | 0.0026 (13) | 0.0127 (9) |
| | 4 | 18.46 | 22.89 | 31.73 | 29.50 | 14.52 |
| | 5 | 0.2391 (15) | 0.0865 (15) | 0.2021 (15) | 0.0056 (13) | 0.1050 (9) |
| Gamma | 1 | 37.78 | 13.59 | 16.85 | 15.78 | 8.82 |
| | 2 | 0.0016 (16) | 0.6292 (16) | 0.9346 (27) | 0.3272 (14) | 0.5489 (10) |
| | 3 | 0.0006 (14) | 0.4807 (14) | 0.8871 (25) | 0.2017 (12) | 0.3574 (8) |
| | 4 | 18.19 | 6.91 | 12.22 | 6.20 | 6.24 |
| | 5 | 0.1982 (14) | 0.9383 (14) | 0.9847 (25) | 0.9056 (12) | 0.6209 (8) |
| Weibull | 1 | 34.92 | 21.73 | 15.03 | 25.35 | 7.79 |
| | 2 | 0.0041 (16) | 0.1522 (16) | 0.9691 (27) | 0.0313 (14) | 0.6495 (10) |
| | 3 | 0.0015 (14) | 0.0844 (14) | 0.9408 (25) | 0.0132 (12) | 0.4544 (8) |
| | 4 | 18.33 | 14.03 | 12.41 | 12.11 | 6.24 |
| | 5 | 0.1922 (14) | 0.4478 (14) | 0.9829 (25) | 0.4372 (12) | 0.6209 (8) |
| Compound Exponential | 1 | 27.44 | 35.09 | 25.40 | 36.52 | 19.44 |
| | 2 | 0.0368 (16) | 0.0039 (16) | 0.5519 (27) | 0.0009 (14) | 0.0350 (10) |
| | 3 | 0.0169 (14) | 0.0014 (14) | 0.4400 (25) | 0.0003 (12) | 0.0127 (8) |
| | 4 | 15.74 | 22.89 | 14.50 | 31.73 | 11.93 |
| | 5 | 0.3293 (14) | 0.0621 (14) | 0.9522 (25) | 0.0015 (12) | 0.1544 (8) |
| Lognormal | 1 | 17.78 | 14.75 | 30.69 | 29.50 | 16.85 |
| | 2 | 0.3367 (16) | 0.5429 (16) | 0.2841 (27) | 0.0089 (14) | 0.0778 (10) |
| | 3 | 0.2168 (14) | 0.3953 (14) | 0.1995 (25) | 0.0033 (12) | 0.0317 (8) |
| | 4 | 10.71 | 9.96 | 16.33 | 18.65 | 7.27 |
| | 5 | 0.7085 (14) | 0.7653 (14) | 0.9046 (25) | 0.0974 (12) | 0.5077 (8) |
| Mixed Exponential | 1 | 33.97 | 35.09 | 29.78 | 31.89 | 20.99 |
| | 2 | 0.0055 (16) | 0.0039 (16) | 0.3243 (27) | 0.0041 (14) | 0.0212 (10) |
| | 3 | 0.0012 (13) | 0.0008 (13) | 0.1923 (24) | 0.0008 (12) | 0.0038 (7) |
| | 4 | 18.46 | 22.89 | 13.07 | 28.54 | 14.52 |
| | 5 | 0.1407 (13) | 0.0430 (13) | 0.9650 (24) | 0.0027 (12) | 0.0427 (8) |
| Generalised Pareto | 1 | 23.63 | 8.65 | 24.03 | 16.10 | 14.00 |
| | 2 | 0.0979 (16) | 0.9271 (16) | 0.6285 (27) | 0.3076 (14) | 0.1730 (10) |
| | 3 | 0.0347 (13) | 0.7989 (13) | 0.4597 (24) | 0.1376 (11) | 0.0512 (7) |
| | 4 | 9.62 | 4.73 | 12.41 | 11.31 | 9.34 |
| | 5 | 0.7244 (13) | 0.9807 (13) | 0.9748 (24) | 0.4178 (11) | 0.2291 (7) |

Table 7. Continued.

| Distribution | (1) | Hospital nursing | In-patient services from specialists | Out-patient services from specialists | Physiother- apy and expedients | Transport- ation services |
|-----------------------------|-----|---------------------|---|--|--------------------------------------|---------------------------------|
| Burr | 1 | 18.19 | 8.21 | 22.01 | 12.74 | 15.55 |
| | 2 | 0.3128 (16) | 0.9422 (16) | 0.7369 (27) | 0.5467 (14) | 0.1132 (10) |
| | 3 | 0.1504 (13) | 0.8294 (13) | 0.5787 (24) | 0.3103 (11) | 0.0295 (7) |
| | 4 | 9.49 | 6.32 | 12.28 | 8.60 | 8.82 |
| | 5 | 0.7352 (13) | 0.9335 (13) | 0.9765 (24) | 0.6592 (11) | 0.2656 (7) |
| Log-Student | 1 | 19.55 | 23.32 | 30.69 | 23.28 | 15.55 |
| | 2 | 0.2411 (16) | 0.1053 (16) | 0.2841 (27) | 0.0559 (14) | 0.1132 (10) |
| | 3 | 0.1070 (13) | 0.0379 (13) | 0.1629 (24) | 0.0162 (11) | 0.0295 (7) |
| | 4 | 9.62 | 9.96 | 15.86 | 18.01 | 7.53 |
| | 5 | 0.7244 (13) | 0.6974 (13) | 0.8993 (24) | 0.0813 (11) | 0.3759 (7) |
| Box-Cox | 1 | 28.39 | 7.49 | 20.83 | 21.20 | 14.00 |
| | 2 | 0.0284 (16) | 0.9627 (16) | 0.7940 (27) | 0.0966 (14) | 0.1730 (10) |
| | 3 | 0.0080 (13) | 0.8727 (13) | 0.6485 (24) | 0.0313 (11) | 0.0512 (7) |
| | 4 | 11.12 | 5.02 | 13.13 | 10.67 | 7.53 |
| | 5 | 0.6008 (13) | 0.9748 (13) | 0.9639 (24) | 0.4713 (11) | 0.3759 (7) |
| Number of obser- vations | | 250 | 234 | 858 | 188 | 85 |
| Number of classes | | 17 | 17 | 28 | 15 | 11 |

- (1) 1. Chi-square value based on maximum likelihood estimates, with
 2. - P-value (number of degrees of freedom equal to k-t-1),
 3. - P-value (number of degrees of freedom equal to k-1).
 4. Chi-square value based on minimum chi-square estimates, with
 5. - P-value (number of degrees of freedom).

Table 8. Parameter estimates of the probability models, with standard errors between brackets.

| Explanatory variables | Hospital nursing | In-patient services from specialists | Out-patient services from specialists |
|--|---------------------|--------------------------------------|---------------------------------------|
| μ | | | |
| constant term | | | |
| $x_1 = 1$ | 7.5114 (0.0812) | 6.3222 (0.0608) | 5.1043 (0.1260) |
| age | | | |
| $x_2 = 1 (\geq 65)$ | 0.6396 (0.1129) | 0.3160 (0.0771) | 0.3915 (0.0541) |
| $x_3 = 1 (40-64)$ | 0.3120 (0.0843) | 0.1651 (0.0595) | 0.3111 (0.0421) |
| sex | | | |
| $x_4 = 1$ (female) | -0.1765 (0.0725) | | |
| working hours / type of job | | | |
| $x_6 = 1$ (no job) | | | 0.2194 (0.1238) |
| $x_7 = 1$ (full time) | | | 0.2514 (0.0743) |
| $x_8 = 1$ (wage earner) | | | 0.3071 (0.1214) |
| $x_9 = 1$ (self- employed) | | | 0.2011 (0.1284) |
| place of residence | | | |
| $x_{11} = 1$ (average sized or large town) | | -0.0773 (0.0650) | |
| $x_{12} = 1$ (dormitory town or small town) | | 0.0829 (0.0650) | |
| insured class | | | |
| $x_{13} = 1$ (class II) | 0.2665 (0.0790) | 0.3914 (0.0602) | |

Table 8. Continued.

| Explanatory variables | Hospital nursing | In-patient services from specialists | Out-patient services from specialists |
|---|---------------------|--------------------------------------|---------------------------------------|
| σ | | | |
| constant term | | | |
| $x_1 = 1$ | 2.8327 (0.0574) | | 1.2933 (0.0216) |
| age | | | |
| $x_2 = 1 (\geq 65)$ | 0.9296 (0.0566) | | |
| $x_3 = 1 (40-64)$ | -0.0430 (0.0475) | | |
| family | | | |
| $x_4 = 1$ (female) | | -0.1793 (0.0263) | |
| working hours | | | |
| $x_6 = 1$ (part-time) | 0.2146 (0.0775) | | |
| $x_7 = 1$ (no job) | 0.0693 (0.0472) | | |
| place of residence | | | |
| $x_{11} = 1$ (average sized or large town) | -0.1161 (0.0524) | | |
| $x_{12} = 1$ (dormitory town or small town) | 0.1903 (0.0519) | | |
| insured class | | | |
| $x_{13} = 1$ (class II) | | 0.1272 (0.0417) | 0.0913 (0.0297) |
| GP insurance | | | |
| $x_{14} = 1$ (yes) | | | |
| Log likelihood | -11804.76 | -9526.53 | -28053.14 |
| Number of observations | 1251 | 1204 | 4093 |

Table 9. P-values with respect to the probability models, for several sets of explanatory variables.

| Omitted / added explanatory variables (*) | Hospital nursing | In-patient services from specialists | Out-patient services from specialists |
|---|-------------------------|---|--|
| <u>Omitted</u> | | | |
| μ | | | |
| age: x_2, x_3 | 0.1138×10^{-6} | 0.1655×10^{-3} | 0.5901×10^{-16} |
| sex: x_4 | 0.1507×10^{-1} | | |
| working hours: x_6, x_7 | | | |
| working hours / type of job: x_6, x_7, x_8, x_9 | | | 0.3802×10^{-3} |
| place of residence: x_{11}, x_{12} | | 0.2839×10^{-1} | |
| insured class: x_{13} | 0.7582×10^{-3} | 0.1291×10^{-9} | |
| σ | | | |
| age: x_2, x_3 | 0.4824×10^{-1} | | |
| family: x_5 | | 0.9648×10^{-10} | |
| working hours: x_6, x_7 | 0.1613×10^{-1} | | |
| place of residence: x_{11}, x_{12} | 0.1028×10^{-2} | | |
| insured class: x_{13} | | 0.1912×10^{-2} | 0.1945×10^{-2} |
| GP insurance: x_{14} | | | |
| <u>Added</u> | | | |
| μ | | | |
| age: x_2, x_3 | | | |
| sex: x_4 | | 0.6979 | 0.8614 |

Table 9. Continued.

| Omitted / added explanatory variables (*) | Hospital nursing | In-patient services from specialists | Out-patient services from specialists |
|--|---------------------|---|--|
| family: x_5 | 0.1150 | 0.8280 | 0.3416 |
| working hours: x_6, x_7 | 0.6259 | 0.4280 | |
| type of job: x_8, x_9, x_{10} | 0.5651 | 0.7496 | |
| place of residence: x_{11}, x_{12} | 0.5167 | | 0.2807 |
| insured class: x_{13} | | | 0.6828 |
| GP insurance: x_{14} | 0.6163 | 0.6665 | 0.0722 |
| σ | | | |
| constant term: x_1 | | 0.7345 | |
| age: x_2, x_3 | | 0.9698 | 0.6214 |
| sex: x_4 | 0.4905 | 0.4114 | 0.2408 |
| family: x_5 | 0.1153 | | 0.6286 |
| working hours: x_6, x_7 | | 0.3839 | |
| type of job: x_8, x_9, x_{10} | 0.5609 | 0.7991 | |
| type of job / working hours x_6, x_7, x_8, x_9 | | | 0.8046 |
| place of residence: x_{11}, x_{12} | | 0.1344 | 0.7130 |
| insured class: x_{13} | 0.0544 | | |
| GP insurance: x_{14} | 0.3890 | 0.3418 | 0.6129 |

*) First, each of the explanatory variables, or group of explanatory variables, from the optimal sets of explanatory variables, given in Table 8 is (are) omitted from this set one by one. Next, each of the explanatory variables, or group of explanatory variables, which do(es) not belong to the optimal set is (are) added to this set one by one.

Table 10. Values of the log likelihood and p-values with respect to the probability models.

| Value log likelihood | Hospital nursing | In-patient services from specialists | Out-patient services from specialists |
|--|--------------------------|---|--|
| Based on the optimal of explanatory variables | -11804.76 | -9526.53 | -28053.14 |
| based on ML estimates of μ and σ | -11860.34 | -9585.20 | -28110.51 |
| p-value | 0.3100×10^{-19} | 0.5851×10^{-23} | 0.9586×10^{-22} |

6. Summary and conclusions

In this paper we considered a statistical model for the costs of medical services of an individual insured person during a year. We analyzed separately the probability of having non-zero expenses for medical services: p , and the sum of the non-zero expenses during a year: Y . For both analyses we introduced a model to trace which factors influence the size of the components. For the empirical part of the study we had at our disposal insurance and claim data of 35,246 privately insured Dutch persons.

In order to analyse the expenses for medical services it was necessary to classify them according to a number of different categories. We distinguished seven categories of medical services. With respect to two of them we had too little data. We therefore restricted our analysis to the following five categories: hospital nursing, in-patient services from specialists, out-patient services from specialists, physiotherapy and expedients and transportation services.

The probability p and the parameters of the probability distributions for Y will, in principle, differ for different persons, depending on personal characteristics and on characteristics of the insurance. We assumed that p and the parameters of the probability distributions are functions of these characteristics.

For the probability of having non-zero expenses for medical services we

applied a logit model. Because the probability of having non-zero expenses for hospital nursing and the probability of having non-zero expenses for in-patient services from specialists are strongly related to each other, we took both categories together as in-patient medical services.

It appeared that for all four categories of medical services "age", "insured class" and "GP insurance" contribute considerably to the explanation of the probability. For out-patient services from specialists "place of residence" also played an important role, whereas for in-patient medical services "working hours" appeared to be a significant explanatory variable.

We cannot derive a probability distribution for the amount of the expenses for medical services theoretically. The choice of such a distribution must therefore be determined on empirical considerations. We considered ten probability distributions. To be able to examine which probability distributions are most suitable for the amount of the expenses for medical services of a certain category, we considered the amounts of the expenses of a selected group of insurants, who are more or less homogeneous with respect to their personal characteristics and the characteristics of their insurance. We decided to analyse the group of male insurants aged 40-64 years who have a class III insurance. For each of the five categories of expenses, ten probability distributions were fitted to the data. To estimate the parameters of the distributions, we applied the method of maximum likelihood. As test criterion we used the chi-square test statistic based on estimated equiprobable classes.

The conclusions can be summarized as follows.

1. The Burr distribution is the only one that behaved satisfactory for all five categories of medical expenses.
2. The gamma and the generalized Pareto distribution did well for four categories. However, it could not satisfactory describe the category "hospital nursing", which is an important category from a financial point of view.
3. The lognormal distributions described three categories quite well, Notice that these three categories are the most important ones from a financial point of view.
4. The Weibull distribution described two categories very well and one category reasonable well.
5. It appeared that the exponential, the compound exponential, the mixed

exponential, the log-t and the Box-Cox distribution were not suitable to describe the amount of expenses for medical services.

In order to trace the factors which influence the parameters of the distributions of the amount of the non-zero expenses for medical services we chose as a first step in the analysis and for simplicity reasons the lognormal distribution. The introduction of explanatory variables for μ as well as for σ resulted in a much higher value of the log likelihood function than the value of the log likelihood function obtained in the case that we did not introduce these explanatory variables

To test whether the contribution of an explanatory variable is significant, we applied a Generalized Likelihood Ratio test, with significance level equal to 0.05. It appeared that

1. with respect to the category "hospital nursing" the variables "age", "sex", "working hours", "place of residence" and "insured class",
 2. with respect to the category "in-patient services from specialists" the variables "age", "family" and "place of residence, and
 3. with respect to the category "out-patient services from specialists" the variables "age", "working hours / type of job" and "insured class"
- contributed significantly to the value of the relevant log likelihood function. We hope to do more research in this field in the coming years.

References.

- Amemiya, T. (1981), "Qualitative response models: a survey", Journal of Economic Literature, Vol. 19, pp. 1483-1536.
- Andersen, R. (1968), A behavioral model of families' use of health services, Center for Health Administration Studies Research Series No. 25, University of Chicago, Chicago.
- Andersen, R., J. Kravitz and O.W. Andersen (1975), Equity in health: empirical analysis in social policy, Ballinger Publ. Cie., Cambridge, Mass.
- Beard, R.E., T. Pentikäinen and E. Pesonen (1977), Risk theory, the stochastic basis of insurance, Second edition, Chapman and Hall, London.
- Chernoff, H. and E.L. Lehmann (1954), "The use of maximum likelihood estimates in χ^2 tests for goodness of fit", The Annals of Mathematical Statistics, Vol. 25, pp. 579-586.

- Dahiya, R.C. and J. Gurland (1972), "Pearson chi-squared test of fit with random intervals", Biometrika, Vol. 59, pp. 147-153.
- Gumbel, E.J. (1943), "On the reliability of the classical χ^2 test", The Annals of Mathematical Statistics, Vol. 14, pp. 253-263.
- Hogg, R.V. and S.A. Klugman (1983), "On the estimation of long tailed skewed distributions with actuarial applications", Journal of Econometrics, Vol. 23, pp. 91-102.
- Mann, H.B. and A. Wald (1942), "On the choice of the number of class intervals in the application of the chi-square test", The Annals of Mathematical Statistics, Vol. 13, pp. 306-317.
- Manning, W.G., C.N. Morris, J.P. Newhouse, L.L. Orr, N. Duan, E.B. Keeler, A. Leibowitz, K.H. Marquis, M.S. Marquis and C.E. Phelps (1981), "A two-part model of the demand for medical care: preliminary results from the health insurance study", in: J. van der Gaag and M. Perlman (eds.), Health, economics, and health economics, North-Holland Publ. Cie., Amsterdam-New York-Oxford.
- McCullagh, P. and J.A. Nelder (1983), Generalised linear models, Chapman and Hall, London/New York.
- Newhouse, J.P., J.E. Rolph, B. Mori and M. Murphy (1980), "The effect of deductibles on the demand for medical care services", Journal of the American Statistical Society, Vol. 75, pp. 525 - 533.
- Van de Ven, W.P.M.M. (1981), "Ziekenfonds- versus particuliere verzekeringen in de gezondheidszorg (I) and (II)", Economisch-Statistische Berichten, Vol. 66, pp. 524-530 and pp. 552-557.
- Van de Ven, W.P.M.M. and J. van de Gaag (1982), "Health as an unobservable, a MIMIC-model of demand for health care", Journal of Health Economics, Vol. 1, pp. 157-183.
- Van de Ven, W.P.M.M., F.A. Nauta, R.C.J.A. van Vliet and F.F.H. Rutten (1980), "Inventarisatie en achtergronden van de consumptieverschillen tussen ziekenfonds en particulier verzekerden", Gezondheid en Samenleving, Vol. 1, pp. 224-254.
- Weber, D.C. (1970), A stochastic model for automobile accident experience, Institute of Statistics, Mimeograph Series No. 651, North Carolina State University, Raleigh, North Carolina.
- Wilks, S.S. (1962), Mathematical Statistics, John Wiley & Sons Inc., New York/London.
- Williams, C.A. (1950), "On the choice of the number and the width of classes for the chi-square test of goodness of fit", Journal of the American Statistical Association, Vol. 45, pp. 77-86.

Appendix

We introduced nine probability distributions in Subsection 5.2. In this Appendix we present the probability density functions (p.d.f.) belonging to them.

One-parameter distribution

The exponential distribution is characterized by

$$(A.1) \quad g(y) = \frac{1}{\beta} \exp\left\{-\frac{y}{\beta}\right\}, \quad y > 0; \quad \beta > 0.$$

Two-parameter distributions

A generalization of the exponential distribution is the gamma distribution. Its p.d.f. is defined as

$$(A.2) \quad g(y) = \frac{y^{\alpha-1}}{\Gamma(\alpha)\beta^{\alpha}} \exp\left\{-\frac{y}{\beta}\right\}, \quad y > 0; \quad \alpha, \beta > 0.$$

If a random variable X is exponentially distributed with parameter β , then it holds that the random variable $Y = X^{1/\tau}$ is Weibull distributed with p.d.f. given by

$$(A.3) \quad g(y) = \frac{\tau}{\beta} y^{\tau-1} \exp\left\{-y^{\tau}/\beta\right\}, \quad y > 0; \quad \tau, \theta > 0.$$

If a random variable Y is conditionally exponential distributed with parameters $\lambda = 1/\beta$, and λ is assumed to be gamma distributed with parameters μ and β , then the unconditional distribution of Y is given by

$$(A.4) \quad g(y) = \frac{\alpha\beta}{(1 + \beta y)^{\alpha+1}}, \quad y > 0; \quad \alpha, \beta > 0,$$

which is the p.d.f of a compound exponential distribution.

A fourth two-parameter distribution is the lognormal distribution. Its p.d.f. is given by

$$(A.5) \quad g(y) = \frac{1}{\sigma y \sqrt{2\pi}} \exp\left\{-\frac{(\log y - \mu)^2}{2\sigma^2}\right\}, \quad y > 0; \quad \sigma > 0.$$

Three-parameter distributions

Another generalization of the exponential distribution is a mixed exponential distribution, defined as a weighted sum of exponential distributions. The three-parameter mixed exponential distribution has the following p.d.f.:

$$(A.6) \quad g(y) = \alpha \frac{1}{\theta_1} e^{-y/\theta_1} + (1-\alpha) \frac{1}{\theta_2} e^{-y/\theta_2}, \quad y > 0; \quad 0 \leq \alpha \leq 1, \quad \theta_1, \theta_2 > 0.$$

Obviously, when α tends to 0 or 1, (A.6) tends to an exponential distribution.

Assume Y is conditionally gamma distributed with parameters $k = \alpha$ and $\theta = 1/\beta$, and θ is gamma distributed with parameters α and $\lambda = 1/\beta$. Then the unconditional p.d.f. of Y is given by

$$(A.7) \quad g(y) = \frac{\Gamma(k+\alpha)}{\Gamma(k) \Gamma(\alpha)} \frac{(y/\lambda)^{k-1}}{\lambda(1 + y/\lambda)^{k+\alpha}}, \quad y > 0; \quad k, \alpha, \lambda > 0.$$

This is the p.d.f. of a generalized Pareto distribution.

Assume Y is conditionally Weibull distributed with parameters τ and $\theta = 1/\beta$, and θ is gamma distributed with parameters α and $\lambda = 1/\beta$, then the unconditional p.d.f. of Y results in

$$(A.8) \quad g(y) = \frac{\alpha \tau y^{\tau-1}}{\lambda(1 + \frac{y^\tau}{\lambda})^{\alpha+1}}, \quad y > 0; \quad \alpha, \tau, \lambda > 0.$$

which is the p.d.f. of a Burr distribution. This distribution can also be derived from (A.7). Let, therefore, X generalized Pareto distributed with parameter $k = 1$, then $Y = X^{1/\tau}$ is Burr distributed with parameters α , τ and λ .

From (A.4) we can derive a distribution which is called the log-t distribution. Let Y be conditionally log normally distributed with mean equal to μ and variance equal to $\tau^2/(2\theta)$. Then its p.d.f. is equal to

$$f(y|\theta) = \frac{\theta^{\frac{1}{2}}}{\tau y \sqrt{\pi}} \exp\left\{-\theta \frac{(\log y - \mu)^2}{\tau^2}\right\}, \quad y > 0; \quad \theta, \tau > 0.$$

Let θ be gamma distributed with p.d.f. equal to

$$h(\theta) = \frac{v^{\frac{1}{2}v} \theta^{\frac{1}{2}v-1}}{\Gamma(\frac{1}{2}v)} \exp\{-v\theta\}, \quad \theta > 0; \quad v > 0.$$

Then the unconditional distribution is given by

$$(A.9) \quad g(y) = \frac{\Gamma\{\frac{1}{2}(v+1)\}}{\sqrt{\pi} \Gamma(\frac{1}{2}v)} \frac{1}{y\tau\sqrt{v}} \left(1 + \frac{(\log y - \mu)^2}{v\tau^2}\right)^{-\frac{1}{2}(v+1)}, \quad y > 0; \quad v, \tau > 0$$

It can be proved that a log-t distribution tends to lognormal distribution, when v tends to infinity.

Finally, we present a family of distributions introduced by Box and Cox (1964). They state: let Y be some continuous random variable; then, for certain distributions of Y, there is a value of λ , such that the random variable

$$\begin{aligned} X &= \frac{Y^\lambda - 1}{\lambda} && \text{for } \lambda \neq 0 \\ &= \log Y && \text{for } \lambda = 0 \end{aligned}$$

is normally distributed. For $\lambda = 0$, Y is lognormally distributed with p.d.f.

given in (A.4). For $\lambda \neq 0$ the p.d.f. of Y is defined as

$$(A.10) \quad g(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2} \left(\frac{y^\lambda - 1}{\lambda} - \mu\right)^2\right\} y^{\lambda-1}, \quad y > 0; \quad \sigma > 0.$$

We will call this distribution a Box-Cox distribution. If $\lambda = 1$ this distribution results in a normal distribution with mean $\mu+1$ and variance σ^2 . It can be shown that if λ tends to $1/3$, this distribution tends to a gamma distribution.

LIST OF REPORTS 1986

- 8600 "Publications of the Econometric Institute Second Half 1985: List of Reprints 415-442, Abstracts of Reports".
- 8601/A T. Kloek, "How can we get rid of dogmatic prior information?", 23 pages.
- 8602/A E.G. Coffman jr, G.S. Lueker and A.H.G Rinnooy Kan, "An introduction to the probabilistic analysis of sequencing and packing heuristics", 66 pages.
- 8603/A A.P.J. Abrahamse, "On the sampling behaviour of the covariability coefficient ζ ", 12 pages.
- 8604/C A.W.J. Kolen, "Interactieve routeplanning van bulktransport: Een praktijktoepassing", 13 pages.
- 8605/A A.H.G. Rinnooy Kan, J.R. de Wit and R.Th. Wijmenga, "Nonorthogonal two-dimensional cutting patterns", 20 pages.
- 8606/A J. Csirik, J.B.G. Frenk, A. Frieze, G. Galambos and A.H.G. Rinnooy Kan, "A probabilistic analysis of the next fit decreasing bin packing heuristic", 9 pages.
- 8607/B R.J. Stroecker and N. Tzanakis, "On certain norm form equations associated with a totally real biquadratic field", 38 pages.
- 8608/A B. Bode and J. Koerts, "The technology of retailing: a further analysis for furnishing firms (II)", 12 pages.
- 8609/A J.B.G. Frenk, M. van Houtveninge and A.H.G. Rinnooy Kan, "Order statistics and the linear assignment problem", 16 pages.
- 8610/B J.F. Kaashoek, "A stochastic formulation of one dimensional pattern formation models", 17 pages.
- 8611/B A.G.Z. Kemna and A.C.F. Vorst, "The value of an option based on an average security value", 14 pages.
- 8612/A A.H.G. Rinnooy Kan and G.T. Timmer, "Global optimization", 47 pages.
- 8613/A A.P.J. Abrahamse and J.Th. Geilenkirchen, "Finite-sample behaviour of logit probability estimators in a real data set", 25 pages.
- 8614/A L. de Haan and S. Resnick, "On regular variation of probability densities", 17 pages
- 8615 "Publications of the Econometric Institute First Half 1986: List of Reprints 443-457, Abstracts of Reports".

- 8616/A W.H.M. van der Hoeven and A.R. Thurik, "Pricing in the hotel and catering sector", 22 pages.
- 8617/A B. Nooteboom, A.J.M. Kleijweg and A.R. Thurik, "Normal costs and demand effects in price setting", 17 pages.
- 8618/A B. Nooteboom, "A behavioral model of diffusion in relation to firm size", 36 pages.
- 8619/A J. Bouman, "Testing nonnested linear hypotheses II: Some invariant exact tests", 185 pages.
- 8620/A A.H.G. Rinnooy Kan, "The future of operations research is bright", 11 pages.
- 8621/A B.S. van der Laan and J. Koerts, "A logit model for the probability of having non-zero expenses for medical services during a year", 22 pages
- 8622/A A.W.J. Kolen, "A polynomial algorithm for the linear ordering problem with weights in product form", 4 pages.
- 8623/A A.H.G. Rinnooy Kan, "An introduction to the analysis of approximation algorithms", 14 pages.
- 8624/A S.R. Wunderink-van Veen and J. van Daal, "The consumption of durable goods in a complete demand system", 34 pages.
- 8625/A H.K. van Dijk, J.P. Hop and A.S. Louter, "An algorithm for the computation of posterior moments and densities using simple importance sampling", 59 pages.
- 8626/A N.L. van der Sar, B.M.S. van Praag and S. Dubnoff, "Evaluation questions and income utility", 19 pages.
- 8627/C G. Renes, A.J.M. Hagenaars and B.M.S. van Praag, "Perceptie en realiteit op de arbeidsmarkt", 18 pages.
- 8628/C B.M.S. van Praag and M.E. Homan, "Lange en korte termijn inkomens-elastiteitscijfers", 16 pages.
- 8629/A R.C.J.A. van Vliet and B.M.S. van Praag, "Health status estimation on the basis of mimic health care models", 32 pages.
- 8630/A K.M. van Hee, B. Huitink and D.K. Leegwater, Portplan, a decision support system for port terminals", 25 pages.
- 8631/A D.K. Leegwater, "Economical effects of delay and acceleration of (un)loading multipurpose ships for stevedore firms", 18 pages.

- 8632/A A.M. Wesselman and B.M.S. van Praag, "Elliptical regression operationalized", 10 pages.
- 8633/C R.C.J.A. van Vliet and E.K.A. van Doorslaer, "De relatie tussen ziekenhuiscapaciteit en -gebruik: een analyse van de gevolgen van aggregatie", 81 pages.
- 8634/B J. Brinkhuis, "Normal integral bases and complex conjugation", 19 pages.
- 8635/A B.S. van der Laan, J. Koerts and J. Reichardt, "A statistical model for the expenses for medical services during a year", 39 pages.

