应用概率统计　第二十八卷
第四期　2012年8月

Chinese Journal of Applied Probability
and Statistics Vol.28 No.4 Aug. 2012

商业医疗保险损失分析: 基于广义线性模型的实证研究 *

仇春涓　　陈　滔

(华东师范大学金融与统计学院, 上海, 200241)　　(西南财经大学保险学院, 成都, 611130)

摘　要

本文使用广义线性模型对商业医疗保险损失进行建模, 并用某商业保险公司的医疗保险赔付数据进行了实证检验, 结果表明, 在影响医疗保险损失的诸多因素中, 住院天数、医院级别、地区、保障档次等都是显著的因素, 而性别和小于60岁以下年龄段内年龄则并不是显著因素, 这些结论给医疗保险的经营和风险控制带来实际的意义.

关键词: 医疗保险, 广义线性模型, 损失分析.

学科分类号: F840.684.

§1. 引　　言

商业医疗保险是健康保险中最重要的内容, 也是社会医疗保障体系的最主要的补充支柱. 医疗保险损失是指医疗消费者通过保险的方式从第三方得到的全部或者部分的补偿. 商业医疗保险损失是参保的患者从商业保险人处得到的医疗费用的补偿. 对商业医疗保险损失分析是商业医疗保险定价基础, 是核保的参考依据, 更是事中和事后医疗保险风险控制的关键.

医疗费用数据往往具有右偏重尾、删失等特点. 在医疗费用中普遍使用的统计方法包括: (1)基于正态分布的方法. (2)数据变换方法, 比如对数变化、Box-Cox变化等(见如Duan et al., 1983; Manning and Mullahy, 2001; O'Hagan and Stevens, 2003等). (3)单分布广义线性模型方法. (4)广义线性模型以外的以偏态分布为基础的参数方法. 比如二参数和三参数的Gamma分布和对数正态(Lognormal)分布被Nixon and Thompson (2004)用来处理医疗服务成本. Lognormal和Weibull分布被Marazzi et al. (1998)用来对医院的住院天数建模. (5)混合参数分布模型方法(Deb and Burgess, 2003). (6)生存分析方法(见如Austin et al., 2003; Basu et al., 2004; Dudley et al., 1993; Lipscomb et al., 1998等). (7)非参数回归和半参数回归方法, 此类回归方法目前在方法学上被大量, 但被应用到医疗费用的分析中较少. (8)针对医疗费用中存在删失数据的方法(见如Lin et al., 1997; Lin, 2003; Lin, 2000; O'Hagan and Stevens, 2004). 然而针对医疗费用右偏重尾特性的专门分析方法(Mihaylova et al., 2010)中尤以数据变换方法和广义线性模型被普遍使用. 在利用广义线性模型的分析

*本文受西南财经大学211工程三期建设项目资助.
本文2011年11月25日收到, 2012年1月11日收到修改稿.

中, Cantoni et al. (2006)分析了一组瑞士的医疗费用数据, 发现住院天数、入院状态(一般入院与急诊入院)、保险类型、年龄、性别、出院状态是影响医疗费用的显著因素. 李致炜等(2008)用广义线性模型对CHNS2006 (2006年中国健康营养调查数据)的医疗健康数据进行了实证研究, 结果表明收入、年龄、性别及社会保障程度对城镇居民基本医疗保险的医疗支出影响显著.

在商业医疗保险损失方面, 早在1974年–1982年美国兰德公司做了著名的兰德健康保险实验(见Newhouse, 1993), 结果表明自付能有效降低人们对医疗服务的利用和医疗费用. 解强等(2009)使用面板数据方法分析了SOA (北美精算师协会)建立的团体医疗保险理赔数据库(Group Medical Insurance Claims Database)的数据, 结果表明, 随着免赔额的增加, 医疗保险的索赔额会逐渐降低, 而每次索赔的超额费用会逐渐增加.

使用广义线性模型分析医疗费用已是比较普遍的一种方法, 但是在商业医疗保险中对医疗保险损失的分析, 由于数据来源的匮乏, 缺少定量的分析. 目前为止更多的文献会定性地讨论影响医疗保险损失的因素以及控制医疗保险风险的方法, 比如Kenneth (2003)从影响可保性因素的角度指出性别、年龄、健康状况、职业、现有的保险等都是影响保险人赔付的重要因素.

本文沿用被广泛使用于医疗费用数据的广义线性模型对商业医疗保险损失进行分析. 详细的对方法的比较, 可以参见(Buntin et al., 2004). 文章以下部分是这样安排的: 第二部分是广义线性模型的简介, 第三部分是本文数据来源的说明, 展示一些数据的描述性结果, 第四部分是实证分析与结果, 第五部分以结论与政策建议结束本文.

§2. 广义线性模型简介

设应变量Y_i, $i = 1, 2, \ldots, n$为医疗保险损失(也即医疗保险赔付金额), 解释变量X_{1i}, $X_{2i}, \ldots, X_{ki}$分别是影响医疗保险的风险因子, 比如可能是性别、年龄、险种、地区、住院天数等. 分析这类问题似乎可以用普通线性模型

$$\mathsf{E}[Y_i|X_i] = \mu_i = X_i^T\beta, \tag{2.1}$$

其中$\beta^T = (\beta_0, \beta_1, \ldots, \beta_k)$, $X_i^T = (1, X_{1i}, X_{2i}, \ldots, X_{ki})$. 但是这个模型的使用需要正态性要求: $Y_i \sim N(\mu_i, \sigma^2)$, 即正态性和同方差性. 然而, 医疗费用以及医疗损失往往是偏态重尾, 也不满足同方差性(Cantoni, 2006; 李致炜等, 2008). 此时一个好的选择是使用广义线性模型GLM (generalized linear models, Nelder and Wedderburn, 1972), 这是一类以指数分布族为基础的模型, 在指数分布族中寻找合适的分布来作为应变量的理论分布, 将其均值的某个函数表示成解释变量的线性函数, 因此不但能够近似偏态重尾的分布, 去掉方差齐性的要求, 还能允许应变量的均值与解释变量之间有一定的非线性关系, 因此对医疗费用数据的建模提供了更为灵活的选择.

广义线性模型的基本思想为: 假设Y是一个随机变量, 其分布来自于某个指数散度族(exponential dispersion family)

$$Y \sim f(y;\theta,\varphi) = \exp\Big[\frac{y\theta - b(\theta)}{a(\varphi)} + c(y,\varphi)\Big], \tag{2.2}$$

其中$\varphi > 0$称为尺度参数, θ称为位置参数, 在某个实数区间内取值. 指数散度族不是一个分布族, 而是一类具有相同形式的分布族的总称, 每个不同的分布族由$c(y,\varphi)$的形式决定, 常见的有二项分布族、Poisson分布族、正态分布族、Gamma分布族以及Pareto分布族等等. 在指数散度族中, 有$b'(\theta) = \mathsf{E}Y \overset{\text{记为}}{=} \mu(\theta)$以及$b''(\theta) = \mathsf{Var}\,(Y)/[a(\varphi)] > 0$. 换句话说, Y的均值是参数θ的增函数$\mu(\theta)$. 如果应变量Y的分布被一些自变量$X = (X_1, X_2, \ldots, X_k)^T$所影响, 使得$Y$的均值会随着某些$X_i$的增加而增加(或者是减小), 这时, 一个较为合理的假定是存在一个适当的单调函数g, 使得$g(\mu) = X^T\beta$, 从而, $\mu = \mathsf{E}[Y|X] = g^{-1}(X^T\beta)$, g称为联接函数(link function). 通常$g(\mu)$当μ变化时, 以$(-\infty,\infty)$为值域. 这相当于规定了 θ与$X^T\beta$之间的一种函数关系, 使得广义线性模型的统计推断可以在极大似然估计的理论框架下得到解决. 特别当Y是非负的, 或者是有偏的分布时, 这样的建模方式比简单的线性模型(2.1)更为合理, 详见Lindsey (1997). 总括来说, 对于一组数据(X_i, Y_i), $i = 1, 2, \ldots, n$, 所谓广义线性模型是指这样的一个分布假定

$$\begin{cases} Y_i \sim f(y_i;\theta_i,\varphi) = \exp\Big[\dfrac{y_i\theta_i - b(\theta_i)}{a_i(\varphi)} + c(y_i,\varphi)\Big] \\ g(\mu_i) = g(b'(\theta_i)) = X_i^T\beta \end{cases}, \qquad i = 1, 2, \ldots, n. \tag{2.3}$$

这使得响应变量的均值可以通过一个联接函数(一般是非线性的)与解释变量的线性形式发生关系. 比如, 如果$Y_i \sim$Gamma分布, 具有密度函数

$$f_{\mu_i,v}(y_i) = \frac{v/\mu_i}{\Gamma(v)}\Big(\frac{vy_i}{\mu_i}\Big)^{v-1}\exp\Big(-\frac{vy_i}{\mu_i}\Big),$$

其中$\mu_i = \mathsf{E}Y_i$, 则

$$f_{\mu_i,v}(y_i) = \exp\Big\{\Big[-\frac{y_i}{\mu_i} - \log(\mu_i)\Big]v + (v-1)\log(y_i) + v\log(v) - \log(\Gamma(v))\Big\}. \tag{2.4}$$

令$\theta_i = -1/\mu_i$, $b(\theta_i) = -\log(-\theta_i)$, $\varphi = 1/v$就得到了形式(2.3).

§3. 数据说明

本文根据2008年某商业保险公司在上海和四川两地推广的一个医疗保险险种的理赔数据, 研究医疗损失对影响因素的响应关系. 我们这里不涉及数据整理的细节, 而只是简单描述数据的结构. 应变量是一份医疗保险合同在一个固定保险期期内的最终赔款额. 影响因素及其解释如下.

(1) 被保险人所在的地区(在模型中用0表示四川地区, 1表示上海地区);

(2) 保障档次: 根据床位费平均每日限额、床位费总限额、药品费平均每日限额、药品费总限额、手术费限额、治疗费用限额等分为三个保障档次, 一档的限额最低, 三档的限额最高;

(3) 被保险人性别(在模型中用0表示男性, 1表示女性);

(4) 年龄: 以岁为单位;

(5) 医院级别: 分为一级、二级、三级(医院功能、设施、技术力量等综合水平越高, 其级别越高);

(6) 住院天数;

(7) 是否放射(在模型中用0表示未放射, 1表示放射);

(8) 是否手术(在模型中用0表示未手术, 1表示手术).

对自变量进行初步的分析, 连续型的变量为年龄和住院天数. 我们可以简单地使用直方图对此进行描述.

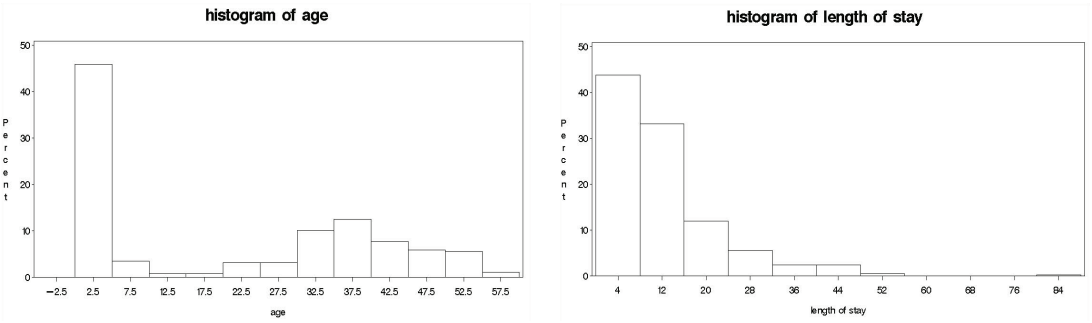

图1 年龄和住院天数的直方图

剩余的自变量都是定性变量, 其相关的分布信息见下表.

表1 定性自变量的描述性统计分析结果

变量	地区		保障档次			性别		医院级别			是否放射		是否手术	
观测	上海	四川	0	1	2	男	女	一级	二级	三级	未放射	放射	未手术	手术
个数	300	77	352	19	6	198	179	12	222	113	267	110	302	75

从图1和表1可以初步看出, 此种医疗保险险种参保的住院患者以儿童居多, 因为此类保险险种针对的是无社保人群, 所以决定了其基本的消费人群儿童会偏多一些. 大部分患者的住院天数不满20天. 上海的参保患者略多于四川, 保障档次主要集中在第一档次(即三类中最低档次), 去一级医院住院就诊的人最少, 因为此类保险保障内容是住院医疗费用, 而一般住院治疗我们都倾向于级别高的医院.

§4. 模　　型

本文使用Gamma分布广义线性模型, 即$Y_i \sim \Gamma(\mu_i, v)$ (见(2.4)式), 联结函数取为$\log(\mu_i) = X_i\beta$. 使用Gamma分布理由如下:

(1) Gamma分布具有方差正比于均值的平方的性质, 有此性质的分布还有Weibull分布等, Blough et al. (1999)和Gilleskie et al. (2004)都观察到医疗费用数据的方差正比于均值的平方. 而Cantoni (2006)和李致炜等(2008), 也曾经使用Gamma分布作为医疗费用的分布, 呈现了较好的结果.

(2) 本文的数据本身首先表明医疗保险损失的分布呈现出方差随均值的增大而增大的趋势(用一般正态线性模型的残差和预测值做散点图可见, 见下图2).

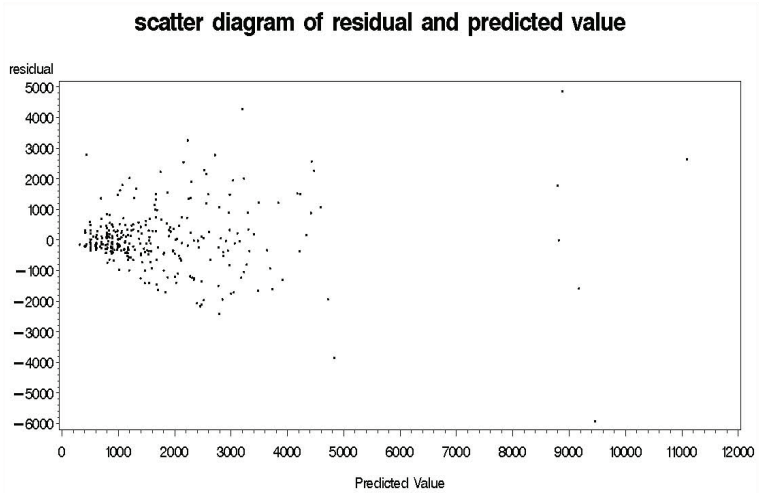

图2　残差与预测值的散点图

(3) 用经典的风险理论来分析的话, 个体在一年的医疗保险损失事实上也是一个复合分布, 即

$$Y_i = \sum_{j=1}^{N_i} X_{ij}, \tag{4.1}$$

其中X_{ij}是第j天的医疗损失, N_i是住院天数. 而往往这样的复合分布我们用Gamma分布去近似会有比较好的效果(Kaas et al., 2001).

鉴于数据中有大量的定性变量, 其具体的三个数据的广义线性模型结构如下

$$\begin{aligned}\log(\mu) \quad = \quad & \beta_0 + \beta_1 \log(\text{住院天数}) + \beta_2 \text{年龄} + \lambda_i^{\text{保障档次}} \\ & + \lambda_j^{\text{地区}} + \lambda_k^{\text{医院级别}} + \lambda_l^{\text{性别}} + \lambda_m^{\text{是否放射}} + \lambda_n^{\text{是否手术}},\end{aligned} \tag{4.2}$$

其中$i = 1, 2, 3$, $j = 0, 1$, $k = 1, 2, 3$, $l = 0, 1$, $m = 0, 1$, $n = 0, 1$.

§5. 实证分析与结果

通过对数据的处理, Pearson Chi-Square值/自由度为0.3575, 小于2, 说明模型拟合的较好. 下表2是详细的参数估计值和95%置性区间.

表2 参数估计和95%置性区间

参数		自由度	估计值	标准差	95%置性区间		卡方值	P值
截距		1	7.9699	0.3167	7.3492	8.5906	633.29	< 0.0001
log(住院天数)		1	0.5875	0.0578	0.4742	0.7007	103.35	< 0.0001
年龄		1	0.0044	0.0024	-0.0003	0.0091	3.33	0.0679
保障档次	1	1	-1.4401	0.2872	-2.0029	-0.8773	25.15	< 0.0001
保障档次	2	1	-1.5183	0.3303	-2.1657	-0.8710	21.13	< 0.0001
保障档次	3	0	0				.	.
地区	0	1	0.1926	0.0950	0.0064	0.3788	4.11	0.0427
地区	1	0					.	.
医院级别	1	1	-0.2288	0.2132	-0.6467	0.1891	1.15	0.2832
医院级别	2	1	-0.1835	0.0757	-0.3319	-0.0350	5.87	0.0154
医院级别	3	0	0				.	.
性别	0	1	0.0670	0.0711	-0.0725	0.2064	0.89	0.3465
性别	1	0	0				.	.
是否放射	0	1	-0.2129	0.0808	-0.3713	-0.0544	6.94	0.0085
是否放射	1	0	0				.	.
是否手术	0	1	-0.4952	0.1014	-0.6940	-0.2964	23.84	< 0.0001
是否手术	1	0	0				.	.
尺度参数		1	2.7779	0.2101	2.3952	3.2218		

假设给定5%的显著性水平, 影响医疗保险损失显著的因素有:

(a) 住院天数对数. 系数是正的, 说明随着住院天数的延长和年龄的增长, 医疗保险的理赔会随之增长. 而且住院天数对数的系数(0.5875)是所有系数中较大的, 住院天数是影响医疗保险赔付的非常重要的一个因素.

(b) 保障档次1和2. 该商业医疗保险是一种补偿型住院医疗费用保险, 针对的是无社保的人群, 但是设置了某些费用类别的限额. 根据床位费的平均每日限额和总限额、药品费的平均每日限额和总限额、护理费用限额、诊疗费用限额、治疗费用限额、检查化验费用限额以及手术费用限额等分为三个档次, 三档的限额高于二档, 二档的限额高于一档. 保障档次1和2的系数显著, 且为负数, 说明保障档次越低, 赔付越低. 进一步数据参数估计结果

显示, 保障档次2的赔付是保障档次3的赔付的$\exp(-1.5183) = 0.219$. 但是保障档次1的赔付却与保障档次2的赔付差不多($\exp(-1.4401)/\exp(-1.5183) = 1.08$). 我们从保险条款上可以看到, 保障档次1与保障档次2的限额之差和保障档次2与保障档次3的限额之差几乎完全相同, 然后最终导致的理赔之差没有产生相同的效果, 我们追溯其三个档次参保人员患者的某些项医疗费用时发现产生的原因主要在与保障档次3的平均每日床位费、手术费和检查化验费特别高, 而其它费用没有产生这么大的差别(详见下表3), 其中主要原因可能是投保第三保障档次的人群的经济水平普遍比较高, 在住院就诊时选择的病房类型、检查化验方式以及手术费用都远远超过了普通病例的费用标准. 同时也说明了从保障档次2提高到保障档次3给风险控制带来了更大的难度.

表3　各保障档次的平均医疗费用

保障档次	观测个数	床位费	每日床位费	药品费	每日药品费	护理费	治疗费	诊疗费	检查化验费	手术费
1	352	1728	142	6980	656	883	2859	705	2425	7747
2	19	2240	119	10764	632	1214	3526	1118	4457	6959
3	6	3526	251	10032	667	815	4369	1040	9629	22938

(c) 地区0, 在地区变量上, 0表示上海, 1表示四川, 代表上海的地区变量系数显著, 且为正的, 说明相同的险种在上海的赔付要比四川的高, 这与地区的医疗服务成本有很大的关系. 进一步从系数可以计算得到, 在其它因素不变的情况下, 上海地区医保患者的赔付是四川地区医保患者的赔付的$\exp(0.1926) = 1.21$倍. 而据官方[1]数据显示, 2008年上海市和四川省两地的人均住院医疗费用分别为10287元和6615元. 商业医疗保险的风险控制措施缩小了地区差异, 但是赔付上的差异产生了地区的不公平性.

(d) 医院级别2, 医院级别2的系数显著, 且为负的, 说明二级医院的参保患者的赔付比三级医院低. 进一步从参数估计值可以计算得到在其它因素保持不变的情形下, 二级医院的赔付是三级医院的赔付的$\exp(-0.1835) = 0.83$倍.

(e) 是否放射0和是否手术0, 这两个变量的系数显著, 且为负的. 在医学上, 需要放射治疗和手术的疾病一般医疗费用相对较高, 所以赔付也相对高.

同时我们也看到, 在5%的显著性水平下不显著的系数有: (a) 医院级别1, 虽然二级医院的赔付明显低于三级医院, 但是一级医院的赔付确并没有显著地低于二级医院, 从表1可以看出, 其主要原因可能在于一般一级医院的空床率比较高, 所以一级医院的例均住院天数较长, 这里存在这明显的医疗资源的浪费和不合理配置. (b) 性别0, 我们普遍认为性别是商业医疗保险定价必须考虑的因素, 因为男女的住院费用有差别, 但是从理赔的结果来看, 男女的住院医疗赔付额却没有显著的差异. (c) 年龄, 一般认为年龄是影响医疗费用的一个

[1]数据来源: 《2009年卫生统计年鉴》, 卫生部网站.

非常重要的因素, 但是在我们这里, 从图1中的左图可以看到, 此险种参保患者的年龄都是在60岁以下, 没涉及到60岁以上的老年人群, 所以年龄的因素并不显著.

为了进一步评估每个影响因素的重要性, 需要做GLM中的type1和type3检验(广义线性模型的type1检验和type3检验与一般线性模型中的type1检验和type3检验类似, type1表示有顺序地检验变量的显著性, type3表示无顺序的. 可参见王丽萍等(2002)).

表4 type1和type3检验

		type1检验		type3检验	
变量	自由度	卡方值	P值	卡方值	P值
log(住院天数)	1	95.59	< 0.0001	88.31	< 0.0001
年龄	1	62.50	< 0.0001	3.37	0.0665
保障档次	2	42.78	< 0.0001	35.02	< 0.0001
地区	1	2.94	0.0867	4.20	0.0404
医院级别	2	15.15	0.0005	6.13	0.0466
性别	1	0.01	0.9336	0.88	0.3473
是否放射	1	10.50	0.0012	7.01	0.0081
是否手术	1	24.18	< 0.0001	24.18	< 0.0001

通过表4可以看出, 在5%的显著性水平下无论是type1检验还是type3检验, 性别和年龄都不是一个显著的风险因子. 而其余的变量都显著, 说明住院天数、年龄、保障档次、地区、医院级别、是否放射、是否手术都是影响商业医疗保险赔付的因素.

§6. 结论与政策建议

本文利用广义线性模型对医疗保险损失进行分析, 发现模型拟合得较好, 并且得出了如下一些有意义的结论与建议.

(1) 性别对医疗保险的赔付无显著影响. 目前国内商业医疗保险的定价和核保都会考虑性别因素, 而社会医疗保险普遍不考虑性别因素, 有些国家为了防止性别歧视, 也取消了商业保险的性别差别费率(比如欧盟规定各保险公司必须从2012年12月21日起, 取消汽车、人寿、医疗等保险业务中存在的性别歧视条款). 因此国内的保险公司也可以大胆地尝试取消性别差别费率.

(2) 对60岁以下的医疗保险参保患者, 年龄不是影响医疗保险赔付的显著的因素. 因此, 商业医疗保险定价和核保其实不必太关注60岁以下的年龄的因素, 应重点关注60岁以上老年人口的医疗费用与医疗保险损失才是关键.

(3) 地区对医疗保险的赔付有显著的影响. 撇开保险的因素, 由于各地区医疗服务成本以及治疗手段等差异, 我国医疗费用的地区差异十分明显. 但是在商业医疗保险中, 相同的险种, 相同的费率在不同的地区的理赔却也存在如此之大的差异, 带来了地区不公平性, 也给商保的经营带来很大的挑战. 商业医疗保险可以参考社会医疗保险, 采用地区不同费率, 但必须考虑与监管一致性.

(4) 医院级别是影响医疗保险赔付的显著因素. 保险公司在风险控制上, 应重点审查三级医院的医疗费用以及一级医院的住院天数, 引导参保人员多去一级医院就诊.

(5) 在同一险种中设置不同的保障档次对控制医疗费用有一定的帮助, 但是我们也看到, 并不是保障档次越低其赔付就越低, 必须有一个合理的尺度.

(6) 从总体来看, 商业医疗保险的保障水平较低, 表1中为无社保的上海人提供的住院医疗保险的人均的赔付才2244.8元, 而同年上海市的平均住院医疗费用已经达到10287元, 可见商业医疗保险的保障水平有待进一步提高.

参 考 文 献

[1] Austin, P.C., Ghali, W.A. and Tu, J.V., A comparison of several regression models for analysing cost of CABG surgery, *Statistics in Medicine*, **22(17)**(2003), 2799–2815.

[2] Basu, A., Manning, W.G. and Mullahy, J., Comparing alternative models: log vs cox proportional hazard? *Health Economics*, **13(8)**(2004), 749–765.

[3] Blough, D.K. and Madden, M.C., Modeling risk using generalized linear models, *Journal of Health Economics*, **18**(1999), 153–171.

[4] Buntin, M.B. and Zaslavsky, A.M., Too much ado about two-part models and transformation? Comparing methods of modeling medical expenditures, *Journal of Health Economics*, **23**(2004), 525–542.

[5] Cantoni, E. and Ronchetti, E., A robust approach for skewed and heavy-tailed outcomes in the analysis of health care expenditures, *Journal of Health Economics*, **25(2)**(2006), 198–213.

[6] Duan, N., Manning, Jr., W.G., Morris, C.N. and Newhouse, J.P., A comparison of alternative models for the demand for medical care, *Journal of Business and Economic Statistics*, **1(2)**(1983), 115–126.

[7] Dudley, R.A., Harrell, Jr., F.E., Smith, L.R., Mark, D.B., Califf, R.M., Pryor, D.B., Glower, D., Lipscomb, J. and Hlatky, M., Comparison of analytic models for estimating the effect of clinical factors on the cost of coronary artery bypass graft surgery, *Journal of Clinical Epidemiology*, **46(3)**(1993), 261–271.

[8] Deb, P. and Burgess, J.A., Quasi-experimental comparison of econometric models for health care expenditures, Department of Economics Working Papers [212], Hunter College, New York, 2003.

[9] Ettner, S.L., Frank, R.G., McGuire, T.G., Newhouse, J.P. and Notman, E.H., Risk adjustment of mental heath and substance abuse payments, *Inquiry*, **35(2)**(1998), 223–239.

[10] Gilleskie, D.B. and Mroz, T.A., A flexible approach for estimating the effect of covariates on health expenditure, *Journal of Health Economics*, **23**(2004), 391–418.

[11] Jørgensen, B., Exponential dispersion models (with discussion), *Journal of the Royal Statistical Society, Series B*, **49(2)**(1987), 127–162.

[12] Kaas, R., Goovaerts, M., Dhaene, J. and Denuit, M., *Modern Actuarial Risk Theory*, Kluwer Academic Publisher, 2001.

[13] Lindsey, J.K., *Applying Generalized Linear Models*, Springer, 1997.

[14] Lin, D.Y., Feuer, E.J., Etzioni, R. and Wax, Y., Estimating medical costs from incomplete follow-up data, *Biometrics*, **53**(1997), 419–434.

[15] Lin, D.Y., Linear regression analysis of censored medical cost data, *Biostatistics*, **1**(2000), 35–47.

[16] Lin, D.Y., Regression analysis of incomplete medical cost data, *Statistics in Medicine*, **22**(2003), 1181–1200.

[17] Lipscomb, J., Ancukiewicz, M., Parmigiani, G., Hasselblad, V., Samsa, G. and Matchar, D.B., Predicting the cost of illness: a comparison of alternative models applied to stroke, *Medical Decision Making*, **18(2 Suppl)**(1998), S39–S56.

[18] Manning, W.G. and Mullahy, J., Estimating log models: to transform or not to transform? *Journal of Health Economics*, **20**(2001), 461–494.

[19] Marazzi, A., Paccaud, F., Ruffieux, C. and Beguin, C., Fitting the distributions of length of stay by parametric models, *Medical Care*, **36(6)**(1998), 915–927.

[20] McCullagh, P. and Nelder, J., *Generalized Linear Models, Second Edition*, Boca Raton: Chapman and Hall/CRC, 1989.

[21] Mihaylova, B., Briggs, A., O'hagan, A. and Thompson, S.G., Review of statistical methods for analysing healthcare resource and costs, *Health Economics*, Published online in Wiley Online Library (wileyonlinelibrary.com), DOI:10.1002/hec.1653, 2010.

[22] Mullahy, J., Much ado about two: reconsidering retransformation and two-part model in health econometrics, *Journal of Health Economics*, **17**(1998), 247–281.

[23] Nelder, J.A. and Wedderburn, R.W.M., Generalized linear models, *Journal of the Royal Statistical Society*, Series A (General), **135(3)**(1972), 370–384.

[24] Nixon, R.M. and Thompson, S.G., Parametric modelling of cost data in medical studies, *Statistcs in Medicine*, **23(8)**(2004), 1311–1331.

[25] O'Hagan, A. and Stevens, J.W., Assessing and comparing costs: how robust are the bootstrap and methods based on asymptotic normality? *Health Economics*, **12(1)**(2003), 33–49.

[26] O'Hagan, A. and Stevens, J., On estimators of medical costs with censored data, *Journal of Health Economics*, **23**(2004), 615–625.

[27] 陈滔, 医疗保险精算和风险控制方法, 西南财经大学出版社, 2002.

[28] 李致炜, 宋世斌, 城镇居民基本医疗保险中的医疗费用分析及预测, 统计与决策, **16**(2008), 72–74.

[29] 解强, 李秀芳, 免赔额与医疗保险索赔额关系分析, 系统管理学报, **18(6)**(2009), 672–675.

[30] 任士泉, 陈峰, 杨树勤, 刘德成, 用MCMC方法研究统筹医疗保险的损失分布模型, 中国卫生事业管理, **2**(2001), 122-124.

[31] 王丽萍, 马林茂, 用sas软件拟合广义线性模型, 中国卫生统计, **19(1)**(2002), 50-53.

[32] Black, Jr., K. and Skipper, Jr., H.D., *Life and Health Insurance* (*Thirteenth Edition*), 孙祁祥, 郑伟等译, 经济科学出版社, 2003.

Risk Factors of Losses of Commercial Medical Insurance: An Empirical Analysis Using Generalized Linear Models

Qiu Chunjuan

(*School of Finance and Statistics, East China Normal University, Shanghai, 200241*)

Chen Tao

(*School of Insurance, Southwestern University of Finance and Economics, Chengdu, 611130*)

The risk factors of commercial medical insurance losses are investigated in this paper. We conduct an empirical analysis by fitting the Gamma generalized linear model to a commercial medical insurance's claims data. The result indicates that among the many candidate risk factors of medical insurance losses, the days of hospital stay, the hospital levels, the area where the insurance business is applied and the insurance level are significant. On the contrary, the gender and age are insignificant. Finally, some suggestions are presented, which we believe to be helpful for future medical insurance's operations and management.

Keywords: Medical insurance, generalized linear model, loss analysis.

AMS Subject Classification: 91B30.