上海对外经贸大学

应用回归分析课程论文

基于线性回归分析方法的医疗保险费用预测

学　　院：　统计与信息学院

专　　业：　应用统计学

学　　号：　17064001　17064016

姓　　名：　朱强强　　曹莉莉

任课教师：　周　平

2019 年 12 月

目 录

摘 要

近年来，由于我国消费者保险意识的不断增强，保险需求愈加旺盛，保费制定的合理性很大程度上影响着消费者的选择。虽然在学术界，对于保费的影响因子已有了不同的理论和研究，但目前还是主要集中在宏观、微观经济因素层面，具体探究消费者实际生活状况与保费之间关联关系的研究还比较少。

线性回归分析方法具有模型简单、预测结果准确、模型解释能力强等特点。本文所用数据来自美国一家保险公司抽样美国不同地区医疗保险消费者的实际数据，我们通过建立多元线性回归模型，来探究消费者的性别、身体指数、所在地区等因素与保费之间的关联关系，并进一步做出保险费的合理预测。在最终建立的回归模型中，拟合优度达到 79.74%，模型检验通过，残差的自相关性和异方差性检验不通过，但正态性检验未通过，由于数据量较大，我们认为这是可以接受的。

针对模型的残差正态性未通过的情况，我们在今后可通过添加相关变量和增加数据的方法尝试重新拟合模型。由于模型的残差图有着明显的集聚情况，我们也可通过将数据进行分类之后再拟合模型。

关键词： 医疗保险，影响因素，回归分析

Abstract

In recent years, due to the increasing awareness of consumer insurance in China, the demand for insurance has become stronger and stronger, and the rationality of premium formulation has greatly affected consumer choice. Although there have been different theories and studies on the influence factors of premiums in academia, at present they are mainly focused on the macro and microeconomic factors, and there are relatively few studies that specifically explore the relationship between consumers' actual living conditions and premiums. .The linear regression analysis method has the characteristics of simple model, accurate prediction results and strong model interpretation ability. The data used in this article is from an actual US insurance company sample of medical insurance consumers in different regions of the United States. We established a multivariate linear regression model to explore the relationship between consumers' gender, body index, region and other factors and premiums. And further make a reasonable forecast of insurance premiums. In the finally established regression model, the goodness of fit reached 79.74%, the model test passed, the autocorrelation and heteroscedasticity tests of the residuals failed, but the normality test failed. Due to the large amount of data, we believe that This is acceptable.

In view of the failure of the model's residual normality, we can try to refit the model in the future by adding relevant variables and adding data. Because the residual plot of the model has obvious aggregation, we can also fit the model by classifying the data.

Keywords: medical insurance influential factors regression analysis

一、 问题与背景

1.1 背景

1980 年我国寿险自恢复运营以来，我国保险业市场发展迅速。保险市场层面，截止 2018 年底，我国保费收入达到 5674 亿美元。保险规模位于世界第二。根据普华永道 2018 年 9 月在京发布的《中国保险消费者白皮书》，中国保险业成长性良好，保险消费者日趋走向成熟。技术层面上，随着大数据技术的不断发展以及人工智能等新兴技术的应用，保险业面临着全新的发展前景。近年来，由于我国消费者保险意识的不断增强，我国人均保费不断上升，保险需求也更加旺盛，另一方面，大数据、互联网金融的技术结合也为定制更合理、更人性化的保险策略提供了新的分析思路。

保险金融研究中，保费的影响因子是一个重要的研究领域，通过研究不同消费者的特征以及与保费的关联，进而如何针对不同消费者、不同保险需求来制定最合理的保费，成为了保险金融业重要的研究课题。在学术界，保费的影响因子已有了大量的理论与实证研究。杜林（2011）认为宏观经济社会的增长（国内生产总值、城乡居民储蓄存款等）对人身险保费的影响并不显著，在不同时期，存在着对保费影响更深的因素，例如保险公司的数量、保险业从业人员的规模等。符洋（2018）基于不同人身保险公司面板数据的实证，认为固定资产率、流动资产周转率等是保费收入的主要影响因素。

1.2 问题

目前学术界的主要研究还是立足于国家和地区整体的保费情况，探究具体的宏观和微观经济变量对保费的影响。落实到具体的保险消费者层面，探究消费者的实际状况与保费之间的相关关系的实证研究相对较少。本文试图通过建立具体的回归模型，来探究消费者的具体情况与保费之间的关联关系，并进一步做出保险费的合理预测。

二. 研究数据

2.1 数据来源

本文选择的数据集为医疗保险数据，数据集来自 kaggle（https://www.kaggle.com/mirichoi0218/insurance）网站，为美国的一家保险公司抽样来自美国不同地区的医疗保险消费者的实际数据，包含消费者的具体医疗保险金额以及其自身实际状况，数据共有 7 个变量，1338 个观测。

2.2 变量描述

本文共有 7 个变量，分别是保费(charges)，年龄(age)，抚养的孩子数量(children)，性别(sex)，体重指数(bmi)，是否吸烟(smoker)，所在地区(region)。其中，保费为响应变量，其余变量为预测变量。在预测变量中，年龄、抚养的孩子数量、体重指数为数值变量，其余预测变量为指示变量。

由于 SAS 的 reg 过程步无法为指示变量进行编码，所以我们在数据预处理阶段为指示变量进行编码。针对“性别”变量，我们设置 0 代表女性，1 代表男性；针对“是否吸烟”变量，我们设置 0 代表不吸烟，1 代表吸烟；针对“所在地区”变量，由于其不为二分名义变量，其包含西北地区(northwest)，东北地区(northeast)，西南地区(southwest)，东南地区(southeast)四个值，我们设定 3 个 0-1 变量(region1，region2，region3)，对应关系如下

	region1	**region2**	**region3**
西北地区(northwest)	0	0	0
东北地区(northeast)	1	0	0
西南地区(southwest)	0	1	0
东南地区(southeast)	0	0	1

三.方法概述

3.1 多元线性回归

在现实问题研究中，因变量的变化往往受几个重要因素的影响，此时就需要用两个或两个以上的影响因素作为自变量来解释因变量的变化，这就是多元回归。当多个自变量与因变量之间是线性关系时，所进行的回归分析就是多元线性回归。设某一因变量 y 受 k 个自变量 $x_1, x_2, ..., x_k$ 的影响，其 n 组观测值为（$y_a, x_{1a}, x_{2a}, ..., x_{ka}$），$a = 1,2,...,n$。那么，多元线性回归模型的结构形式为：

$$y_a = \beta_0 + \beta_1 x_{1a} + \beta_2 x_{2a} + ... + \beta_k x_{ka} + \varepsilon_a$$

$\beta_0, \beta_1, ..., \beta_k$ 为待定参数；ε_a 为随机变量。

如果 $b_0, b_1, ..., b_k$ 分别为 $\beta_0, \beta_1, \beta_2 ..., \beta_k$ 的拟合值，则回归方程为

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + ... + b_k x_k$$

b_0 为常数；$b_1, b_2, ..., b_k$ 称为偏回归系数。

在问题中，参数的值与误差的方差不是已知的，必须通过样本数据进行估计。 如果模型满足条件：①$E\varepsilon = 0$； ②$Var(\varepsilon) = \sigma^2 I$；③$X_1, X_2, ..., X_k$ 互不相关，则称模型为普通线性模型。如果模型的随机误差服从正态分布，即 $\varepsilon \sim N(0, \sigma^2 I)$，则称模型为正态线性回归模型。

3.2 回归系数的最小二乘估计

最小二乘法可以用于估计方程中的回归系数。在正态假定下，如果 X 是列满秩的，则普通线性回归模型的参数最小二乘估计为：$(x^T x)^{-1} x^T y$，于是 y 的估计值为：$\hat{y} = x\hat{\beta}$，记残差向量为 $e = y - \hat{y} = y - X\hat{\beta}$，则随机误差 σ^2 的最小二乘估计为：

$\hat{\sigma}^2 = \dfrac{e^T e}{n-k-1}$，得到回归模型参数的估计值。

四. 建立模型

4.1 多元线性模型

本文的目标是建立一个线性回归模型来刻画预测变量（年龄、性别、体重指数、子女数量、是否吸烟、所处地区）和响应变量（保费）之间的关系。首先我们利用所有已知的预测变量来构建一个多元回归模型，构建的模型如下：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \varepsilon$$

其中，y代表响应变量，$x_i, i = 1,2,\cdots,8$依次代表预测变量“年龄”、“身体指数”、“抚养孩子的数量”、“是否吸烟”、“性别”以及表示地区的三个虚拟变量。

方差分析

源	自由度	平方和	均方	F 值	Pr > F
模型	8	1.472347E11	18404336091	500.81	<.0001
误差	1329	48839532844	36749084		
校正合计	1337	1.960742E11			

参数估计

变量	自由度	参数估计	标准误差	t 值	Pr > |t|	方差膨胀
Intercept	1	-12292	988.19590	-12.44	<.0001	0
age	1	256.85635	11.89885	21.59	<.0001	1.01682
bmi	1	339.19345	28.59947	11.86	<.0001	1.10663
children	1	475.50055	137.80409	3.45	0.0006	1.00401
smoker_new	1	23849	413.15335	57.72	<.0001	1.01207
sex_new	1	-131.31436	332.94544	-0.39	0.6933	1.00890
region1	1	352.96390	476.27579	0.74	0.4588	1.51564
region2	1	-607.08709	477.20391	-1.27	0.2035	1.52475
region3	1	-682.05815	478.95916	-1.42	0.1547	1.65407

图 4-1 模型拟合图

拟合的多元线性回归模型的 p 值小于 0.0001，说明该模型是显著的，但是“性别”以及所在地区的三个虚拟变量均未通过系数 t 检验，接受系数为 0 的原假设，模型需要进一步优化。

4.2 逐步回归建立新的多元回归模型

为了优化上述多元回归模型，我们采用逐步回归法将变量依次引入回归模型。逐步回归挑选出来的变量为“是否吸烟”、“年龄”、“身体指数”、“抚养孩子的数

量”以及“所在地区”的第一个虚拟变量“region1”，我们利用这几个变量拟合新的多元线性模型。

方差分析					
源	自由度	平方和	均方	F 值	Pr > F
模型	5	1.471399E11	29427984847	801.03	<.0001
误差	1332	48934297335	36737460		
校正合计	1337	1.960742E11			

图 4-2 逐步回归后的方差分析图

模型 F 检验的 p 值小于 0.0001，说明模型是显著的。为了验证模型满足所有的假设，需要进行四个方面的检验：多重共线性检验、残差的自相关性检验、异方差性检验和正态性检验。

参数估计						
变量	自由度	参数估计	标准误差	t 值	Pr > |t|	方差膨胀
Intercept	1	-12512	963.37491	-12.99	<.0001	0
age	1	257.40583	11.88547	21.66	<.0001	1.01486
bmi	1	329.46287	27.61628	11.93	<.0001	1.03218
children	1	479.51419	137.67426	3.48	0.0005	1.00244
smoker_new	1	23808	410.77277	57.96	<.0001	1.00076
region1	1	773.94622	390.70938	1.98	0.0478	1.02029

图 4-3 逐步回归后的参数估计图

我们由方差因子扩大法判断模型变量之间的多重共线性情况，我们由如图可知所有的 VIF 值均小于 10，所以我们认为该模型的变量之间不存在线性相关性。

第一和第二矩指定的检验		
自由度	卡方	Pr > 卡方
18	181.23	<.0001

Durbin-Watson D	2.083
观测数	1338
第一阶自相关	-0.043

图 4-4 自相关检验的 DW 值

我们使用 SAS 中 spec 命令对模型的残差做异方差性检验，检验的 p 值小于 0.0001，拒绝模型的残差具有异方差性的原假设，即该模型残差不具有异方差性。

Durbin-Watson 检验的统计量为 2.083。一般认为，DW 检验统计量约等于 0 表示残差中存在正自相关性，约等于 4 表示残差中存在负自相关性，约等于 2 表示残差相互独立，所以我们有充足的理由相信该模型残差不具有自相关性。

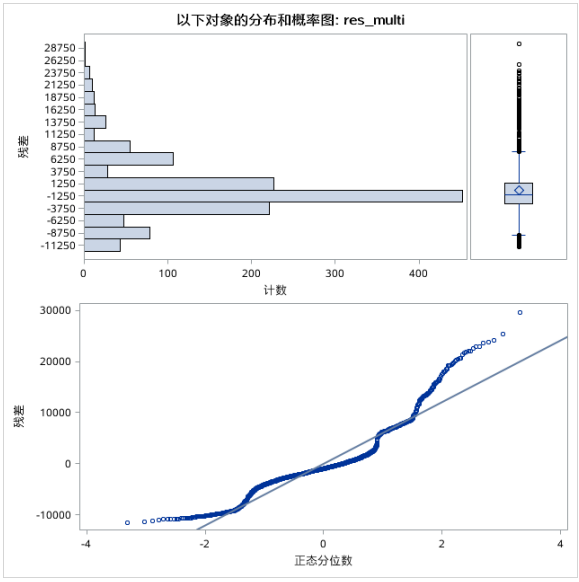

图 4-5 正态分布图

我们对模型的残差做正态性检验，由于数据量小于 5000，所以我们做 Shapiro-Wilk 正态性检验，发现检验统计量的 p 值小于 0.0001，拒绝残差服从正态分布的原假设，残差的正态性检验未通过。

4.3 删除异常值点后构建新的多元回归模型

原数组中的异常值点往往会对模型造成较大的影响，所以删除模型汇总的异常值点是优化模型十分重要的一步。我们给定第i个观测值的学生化残差，即为SRE_i

$$SRE_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

若$|SRE_i| > 3$的观测即判定为异常值。同时，我们也根据库克距离，构建删除异常值的回归模型。

拟合出来的模型为$y = -13268 + 264.81x_1 + 332.80x_2 + 433.94x_3 + 23696x_4 + 751.15x_6$。

方差分析

源	自由度	平方和	均方	F 值	Pr > F
模型	5	1.427926E11	28558521433	1032.41	<.0001
误差	1305	36098997006	27662067		
校正合计	1310	1.788916E11			

图 4-6 剔除异常值后的方差分析图

拟合出来的模型假设检验通过，模型是显著的。

参数估计						
变量	自由度	参数估计	标准误差	t 值	Pr > |t|	方差膨胀
Intercept	1	-13268	847.09777	-15.66	<.0001	0
age	1	264.80865	10.42204	25.41	<.0001	1.01399
bmi	1	332.80079	24.20789	13.75	<.0001	1.03084
children	1	433.94437	120.38681	3.60	0.0003	1.00229
smoker_new	1	23694	360.38764	65.75	<.0001	1.00109
region1	1	751.15393	342.64524	2.19	0.0285	1.02011

图 4-6 剔除异常值后的方差分析图

所有系数均通过假设检验，且 VIF 值均小于 10，不存在方差共线性。

第一和第二矩指定的检验		
自由度	卡方	Pr > 卡方
20	306.21	<.0001

Durbin-Watson D	2.032
观测数	1311
第一阶自相关	-0.017

图 4-7 剔除异常值后的自相关检验的 DW 值

异方差检验的 p 值小于 0.0001，拒绝模型残差存在异方差的原假设。且 DW 检验统计量的值约等于 2，说明模型残差也不存在自相关性。

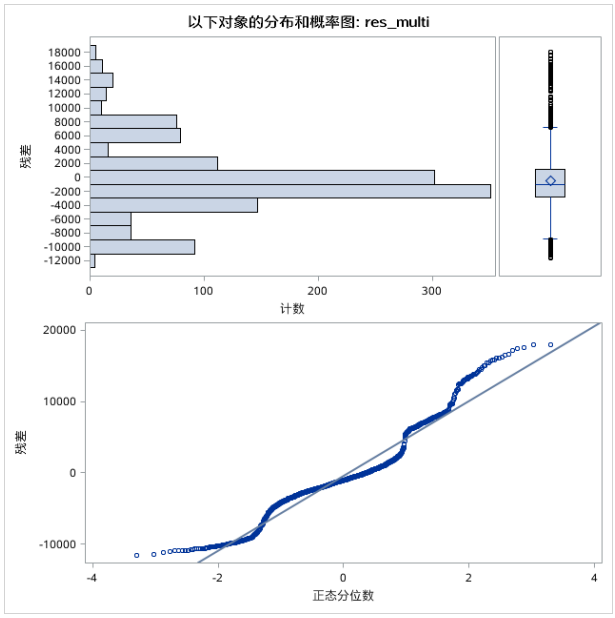

图 4-8 剔除异常值后的正态分布图

在对模型的残差进行 Shapiro-Wilk 检验时，检验的 p 值小于 0.0001，说明模型的残差不满足正态性假设。但由于本文中的样本数量较大，且模型的残差已通过异方差性检验和自相关性检验，我们即使该模型的残差并不符合正态分布，但仍然是可以接受的。

4.4 模型比较

本文一共建立了三个多元线性回归模型，第一个模型是利用所有已知的变量进行拟合，第二个模型我们使用逐步回归方法挑选出最优变量子集来拟合模型，最后一个模型是在第二个模型的技术上利用学生化残差和库克距离删除异常值点后的数据来拟合模型。我们选用拟合优度(R^2)、调整后的拟合优度($Adj-R^2$)以及均方根误差(RMSE)来综合评价我们的模型，结果如下表所示。

Model	$\boldsymbol{R^2}$	$\boldsymbol{Adj-R^2}$	**RMSE**
Model 1	0.7579	0.7494	6062.10
Model 2	0.7504	0.7495	6061.04
Model 3	0.7982	0.7974	5259.47

表 4-1 模型参数对比

从上表的结果来看，模型3的拟合优度和调整后的拟合优度是最高的，且均方误差是最小的，说明模型三最优。然而模型一和模型三这三个评价指标的结果十分相似，说明模型二进行逐步回归删除“性别”以及“地区”中的“region1”和“region2”这两个虚拟变量对模型结果的影响并不是很大。

五.结论

通过对美国保险公司的医疗保险保费数据的实证分析，我们发现性别对保费的影响上不会起到显著的作用。而保险受益人的年龄、身体指数、吸烟状况、抚养子女数量以及所在地区均会对保险费用造成显著的影响。基于模型三这样一个多元线性回归模型，可以为保险公司在保费定价方面提供一个实践参考。

六.不足与展望

模型拟合出来效果良好，自相关性、异方差性、多重共线性检验均通过；但残差的正态性检验并未通过，因此使用本文中的模型预测保费还存在一定的风险，需要进一步扩大样本容量，或者检查该模型未考虑到的其他变量来进一步优化预测模型。

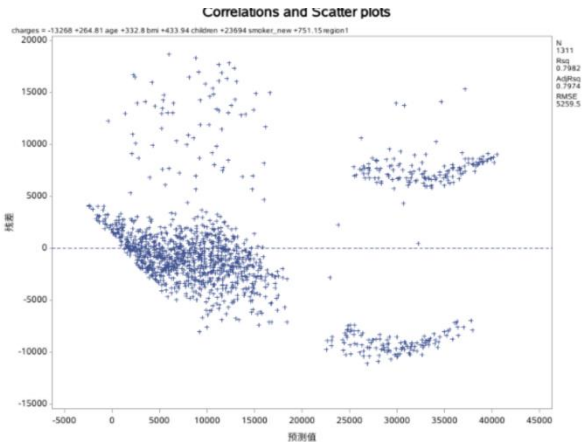

图 6-1 模型残差图

从模型的残差图中我们看出残差有着明显集群，可以分为三类。未来我们可以根据残差的分类将数据进行分类，在数据中添加新的指示变量重新拟合模型，并与本文中的最优模型进行比较，观察新的模型中残差是否通过所有检验。

参考文献

1.何晓群.应用回归分析（第四版）[M]. 北京:中国人民大学出版社, 2015.

2.刘荣.SAS 统计分析与应用实例[M]. 北京:电子工业出版社, 2013.

3.李晓霞.多元线性回归对养老保险基金的预测分析研究[J].商,2015(27):72.

附 录

附录 1. 原始数据

部分数据所附如下，完整数据请见压缩包中 insurance.csv 文件。

age	sex	bmi	children	smoker	region	charges
19	female	27.9	0	yes	southwest	16884.92
18	male	33.77	1	no	southeast	1725.552
28	male	33	3	no	southeast	4449.462
33	male	22.705	0	no	northwest	21984.47
32	male	28.88	0	no	northwest	3866.855
31	female	25.74	0	no	southeast	3756.622
46	female	33.44	1	no	southeast	8240.59
37	female	27.74	3	no	northwest	7281.506
37	male	29.83	2	no	northeast	6406.411
60	female	25.84	0	no	northwest	28923.14

附录 2. SAS 代码

```
proc import datafile='E:\SAS\Final Project\insurance.csv' out=file_data
replace;
    getnames=yes;
run;
proc print data=file_data;
run;
/*为指示变量编码*/
data file_data_new;
    set file_data;
    select;
        when (sex="female") sex_new=0;
        when (sex="male") sex_new=1;
    end;
run;
data file_data_new;
    set file_data_new;
    select;
        when (smoker="no") smoker_new=0;
        when (smoker="yes") smoker_new=1;
    end;
run;
data file_data_new;
    set file_data_new;
    select;
        when (region="northwest") region1=0;
        when (region="northeast") region1=1;
        when (region="southwest") region1=0;
```

```
        when (region="southeast") region1=0;
    end;
run;
data file_data_new;
    set file_data_new;
    select;
        when (region="northwest") region2=0;
        when (region="northeast") region2=0;
        when (region="southwest") region2=1;
        when (region="southeast") region2=0;
    end;
run;
data file_data_new;
    set file_data_new;
    select;
        when (region="northwest") region3=0;
        when (region="northeast") region3=0;
        when (region="southwest") region3=0;
        when (region="southeast") region3=1;
    end;
run;
proc print data=file_data_new;
run;
proc reg data=file_data_new outest=model1_test;
    model charges=age bmi children smoker_new sex_new region1 region2
region3/r corrb vif collin dw spec;
    output out=res_data1 r=res_multi student=stu_res press=press_res
rstudent=rstu_res h=h;
    plot r.*p.;
run;
proc univariate data=res_data1 normal plot;
    var res_multi;
run;
proc print data=model1_test;
run;
proc reg data=file_data_new;
    model charges=age bmi children smoker_new sex_new region1 region2
region3/selection=stepwise;
    plot r.*p.;
run;
proc reg data=file_data_new outest=model2_test;
    model charges=age bmi children smoker_new region1/r corrb vif collin dw
spec;
    output out=res_data2 r=res_multi student=stu_res press=press_res
rstudent=rstu_res h=h;
    plot r.*p.;
run;
proc univariate data=res_data2 normal plot;
    var res_multi;
run;
proc print data=model2_test;
run;
data res_data_model;
    set res_data2;
    if (stu_res>3 or stu_res<-3) then delete;
    if (cookd>0.003) then delete;
```

```
run;
proc print data=res_data_model;
run;
proc reg data=res_data_model outest=model3_test;
    model charges=age bmi children smoker_new region1/r corrb vif collin dw
spec;
    output out=res_data3 r=res_multi;
    plot r.*p.;
run;
proc univariate data=res_data3 normal plot;
    var res_multi;
run;
proc print data=model3_test;
run;
```