数据分析作业 2

——医疗数据分析

16337053 信息安全 杜锦文

环境及工具

SAS STUDIO UNIVERSITY EDITION

数据分析实例

1. 线性回归分析

假设误差服从分布 N(0,sigma2)，建立个人医疗费用和 3 个定量变量之间的线性回归方程并研究相应的统计推断问题。

用前 1333 条数据进行线性回归拟合。

用最后 5 条数据进行测试。预测他的个人医疗费用，并给出置信度为 95%的置信区间。

分析思路

导入文件；

利用SAS的reg过程对模型charges=age bmi children进行分析；

选项cli进行预测，计算置信区间，设置显著性水平为0.05。

CODE

```
PROC IMPORT DATAFILE='/folders/myfolders/data.csv' REPLACE
    DBMS=CSV
    OUT=data0;
    GETNAMES=YES;
RUN;

proc reg data=data0;
model charges=age bmi children/cli alpha=0.05;
```

结果详情见文件‘\result\1.pdf’

SAS 结果

参数估计					
变量	自由度	参数估计	标准误差	t 值	Pr > |t|
Intercept	1	-6872.96706	1761.15212	-3.90	<.0001
age	1	237.74407	22.38976	10.62	<.0001
bmi	1	333.74999	51.40407	6.49	<.0001
children	1	546.27972	259.03701	2.11	0.0351

REG 过程
模型: MODEL1
因变量: charges

输出统计量						
观测	因变量	预测值	标准误差均值预测	95% 置信预测		残差
1334	10601	16997	621.3538	-5346	39340	-6396
1335	2206	8004	632.5175	-14340	30348	-5798
1336	1630	9641	723.7985	-12714	31996	-8011
1337	2008	6691	609.0028	-15650	29033	-4683
1338	29141	17377	662.4615	-4970	39725	11764

SAS 分析

如图，可以得到线性拟合方程 charges=237.74age+333.75bmi+546.28children-6872.97

2. 方差分析

1. 假设个人医疗费用服从方差分析模型，比较不同性别对个人医疗费用是否有显著（显著水平为 0.05）差异。

分析思路

导入文件；

利用SAS的anova过程对模型charges=sex进行分析可以得到结果；

Class选项设置非参数单因素为sex；

means语句计算主效应sex不同水平所对应的因变量均值，设置显著性水平为0.05。

CODE

```
PROC IMPORT DATAFILE='/folders/myfolders/data.csv' REPLACE
    DBMS=CSV
    OUT=data0;
    GETNAMES=YES;
RUN;

proc anova data=data0;
class sex;
model charges=sex;
means sex/alpha=0.05;
run;
```

SAS 结果

ANOVA 过程

因变量: charges

源	自由度	平方和	均方	F 值	Pr > F
模型	1	643590180.13	643590180.13	4.40	0.0361
误差	1336	195430631388	146280412.72		
校正合计	1337	196074221568			

R 方	变异系数	均方根误差	charges 均值
0.003282	91.13986	12094.64	13270.42

源	自由度	Anova SS	均方	F 值	Pr > F
sex	1	643590180.1	643590180.1	4.40	0.0361

结果详情见文件‘\result\2a.pdf’

SAS 分析

可以看到 sex 的 p 值为 0.0361，小于 0.05，拒绝原假设，认为不同性别对个人医疗费用有显著差异。

2. 利用方差分析知识（两因素等重复试验下），假设个人医疗费用服从两因素的方差分析模型，对性别、是否吸烟两个因素，对方差进行分析（显著水平为 0.05）。

分析思路

导入文件；

利用SAS的glm过程对模型charges=sex|smoker进行分析可以得到结果；

Sex|smoker是指对sex、smoker和sex*smoker进行分析；

Class选项设置非参数因素为sex smoker；

means语句计算主效应sex|smoker不同水平所对应的因变量均值，SNK为SNK测验，设置显著性水平为0.05。

SAS 结果

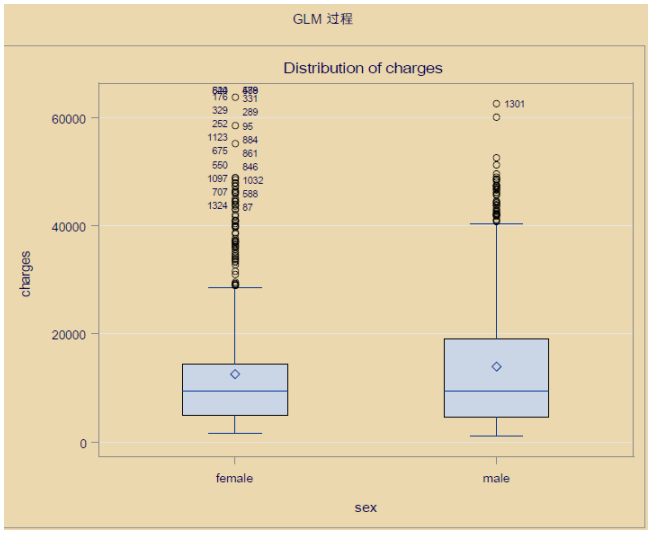
源	自由度	I 型 SS	均方	F 值	Pr > F
sex	1	643590180.13	643590180.13	11.59	0.0007
smoker	1	120877734754	120877734754	2177.28	<.0001
sex*smoker	1	492339740.81	492339740.81	8.87	0.0030

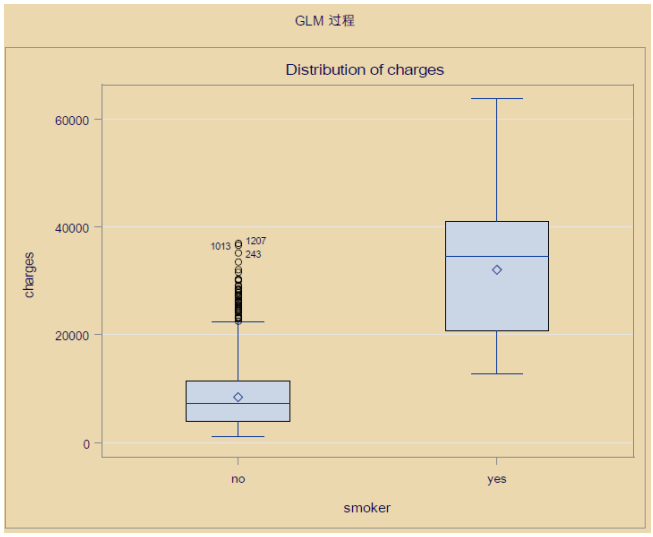
源	自由度	III 型 SS	均方	F 值	Pr > F
sex	1	151971572.34	151971572.34	2.74	0.0983
smoker	1	117186564802	117186564802	2110.80	<.0001
sex*smoker	1	492339740.81	492339740.81	8.87	0.0030

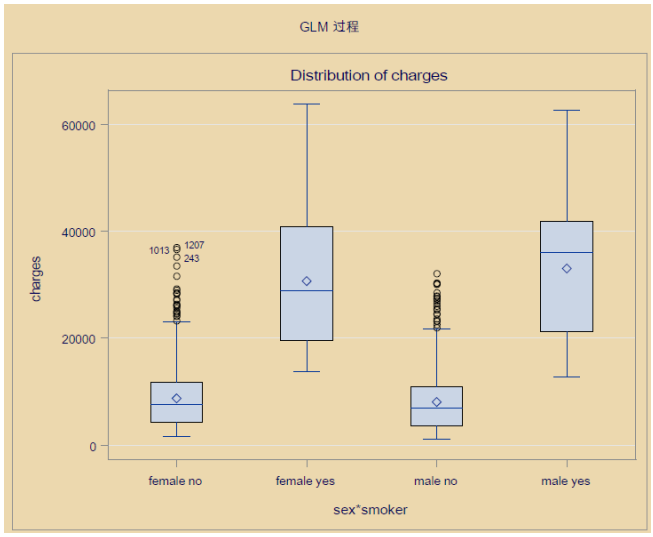
SAS 分析

从图得到 smoker 的 p 值均小于 0.05，sex 的 p 值存在大于 0.05 的现象，则认为是否吸烟对个人医疗费用有显著影响，而性别不是。

更明显的关系我们可以从下面的箱线图中看到：







SAS 分析

很明显，可以说性别对个人医疗支出没有显著影响，而是否吸烟对个人医疗支出有显著影响。