

medical_cost_predict

Rngxin Feng

12/11/2020

```
knitr::opts_chunk$set(echo = FALSE)
```

背景

健康保险在医疗健康系统中发挥着重要的作用，研究健康保险费用支出的相关性以及对影响因素间的模型预测，对于合理设置保险金额，调配保险资源的有效使用，具有重要的意义。

一、可以了解用户在医疗费用支出上的最大、最小值和平均支出费用这些人群分别有什么特点。

二、医疗费用支出与那些因素有关，那些是主要决定因素，如何才能精确的对待不同的人群设定保险额度，以达到资源的最优运用。

三、有没有模型可以反映出大部分用户在健康保险费用支出上的关系，模型的表现如何，是否有大量的异常数据和立群值。

四、依据分析过程，提出相关建议。

目录：

一、加载包并读取数据

- 1.1加载包
- 1.2读取数据
- 1.3变量注释

二、数据清洗

- 2.1检查缺失值
- 2.2数据概览
- 2.3将sex、smoker、region变量转换为因子，并保留在insu_df数据框

三、变量统计以及可视化

- 3.1insurance分布图
- 3.2年龄分布图
- 3.3性别分布图
- 3.4bmi因素分布图
- 3.5儿童个数分布图
- 3.6smoker因素分布图
- 3.7地区因素分布图

四、相关性探索

五、建立模型

- 5.1多元线性模型
- 5.2多重线性
 - 5.2.1以smoker为分组对age变量做多重线性模型的可视化
 - 5.2.2以smoker为分组对bmi变量的多重线性模型可视化

六、结论

一、加载包并读取数据

1.1 加载包

1.2 读取数据

```
## # A tibble: 1,338 x 7
##   age sex    bmi children smoker region  charges
##   <dbl> <chr> <dbl>    <dbl> <chr>  <chr>    <dbl>
## 1    19 female  27.9        0 yes   southwest 16885.
## 2    18 male   33.8        1 no    southeast 1726.
## 3    28 male   33         3 no    southeast 4449.
## 4    33 male   22.7        0 no    northwest 21984.
## 5    32 male   28.9        0 no    northwest 3867.
## 6    31 female 25.7        0 no    southeast 3757.
## 7    46 female 33.4        1 no    southeast 8241.
## 8    37 female 27.7        3 no    northwest 7282.
## 9    37 male   29.8        2 no    northeast 6406.
## 10   60 female 25.8        0 no    northwest 28923.
## # ... with 1,328 more rows
```

1.3 变量注释

序号	变量	注释
1	age	年龄
2	sex	性别
3	bmi	身体质量指数，成人标准值（18.5-23.9），算法： $\text{kg}/(\text{m}^2)$
4	children	小孩数量
5	smoker	是否吸烟
6	region	地区
7	charges	投保费用

二、数据清洗

2.1 检查缺失值

```
## # A tibble: 1 x 7
##   age sex    bmi children smoker region charges
##   <int> <int> <int>    <int> <int> <int>    <int>
## 1     0     0     0        0     0     0        0
```

各个变量均没有缺失值。

2.2数据概览

```
##          age          sex          bmi          children
##  Min.      :18.00   Length:1338   Min.      :15.96   Min.      :0.000
##  1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000
##  Median :39.00   Mode  :character   Median :30.40   Median :1.000
##  Mean    :39.21                      Mean    :30.66   Mean    :1.095
##  3rd Qu.:51.00                      3rd Qu.:34.69   3rd Qu.:2.000
##  Max.     :64.00                      Max.     :53.13   Max.     :5.000
##          smoker          region          charges
##  Length:1338   Length:1338   Min.      : 1122
##  Class :character   Class :character   1st Qu.: 4740
##  Mode  :character   Mode  :character   Median : 9382
##                                     Mean    :13270
##                                     3rd Qu.:16640
##                                     Max.     :63770
```

数据集有1338行（观测值），7个变量，3个字符型向量，4个数字型向量。

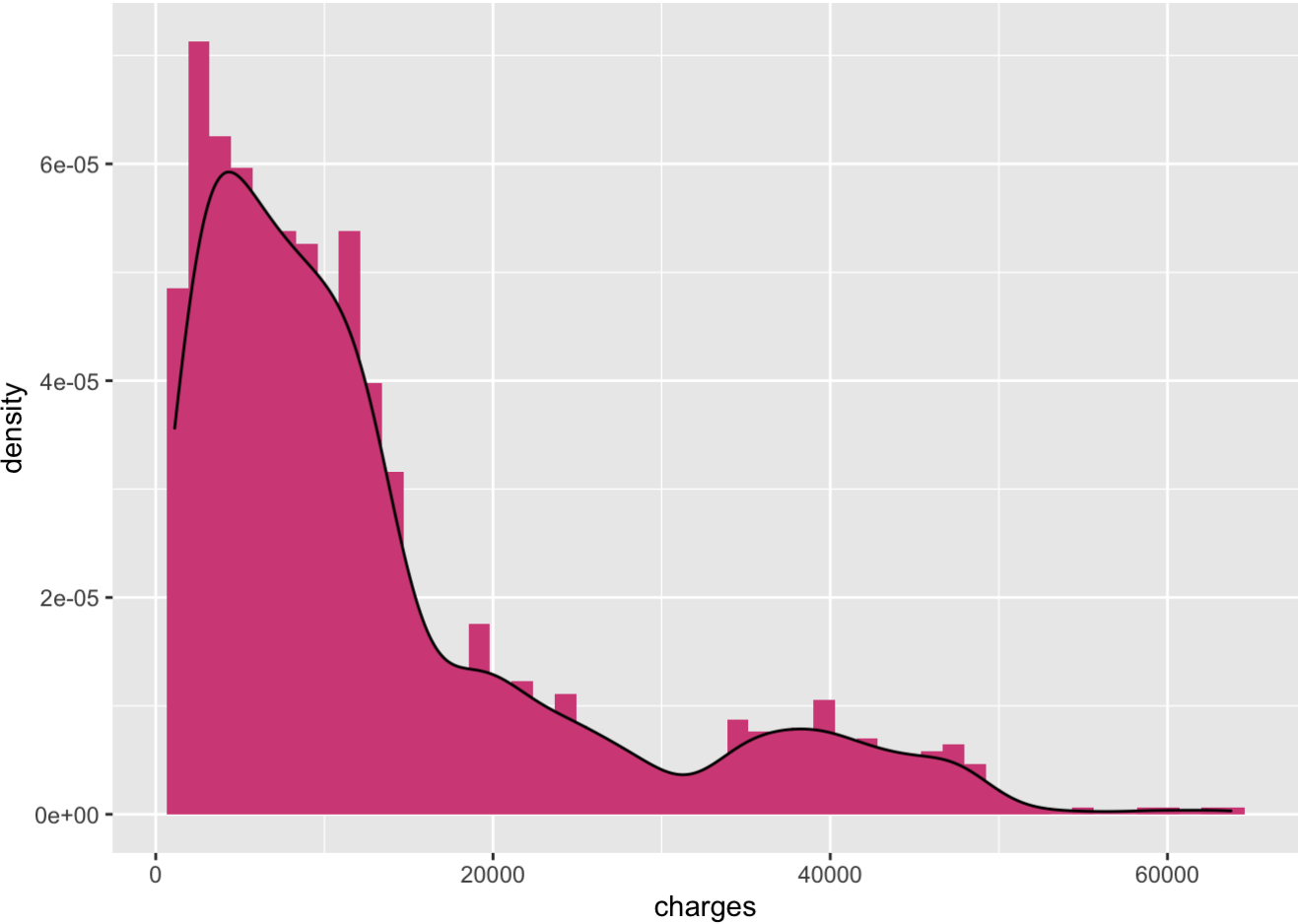
该数据集主要是健康保险费用支出收集的相关数据，自变量包括用户的年龄、性别、身体质量指数、小孩数量、是否吸烟、地区。

2.3将sex、smoker、region变量转换为因子，并保留在insu_df数据框

```
## # A tibble: 1,338 x 7
##   age sex    bmi children smoker region    charges
##   <dbl> <fct> <dbl> <fct>    <fct> <fct>    <dbl>
## 1    19 female  27.9 0        yes    southwest 16885.
## 2    18 male   33.8 1        no     southeast 1726.
## 3    28 male   33   3        no     southeast 4449.
## 4    33 male   22.7 0        no     northwest 21984.
## 5    32 male   28.9 0        no     northwest 3867.
## 6    31 female 25.7 0        no     southeast 3757.
## 7    46 female 33.4 1        no     southeast 8241.
## 8    37 female 27.7 3        no     northwest 7282.
## 9    37 male   29.8 2        no     northeast 6406.
## 10   60 female 25.8 0        no     northwest 28923.
## # ... with 1,328 more rows
```

三、变量统计以及可视化

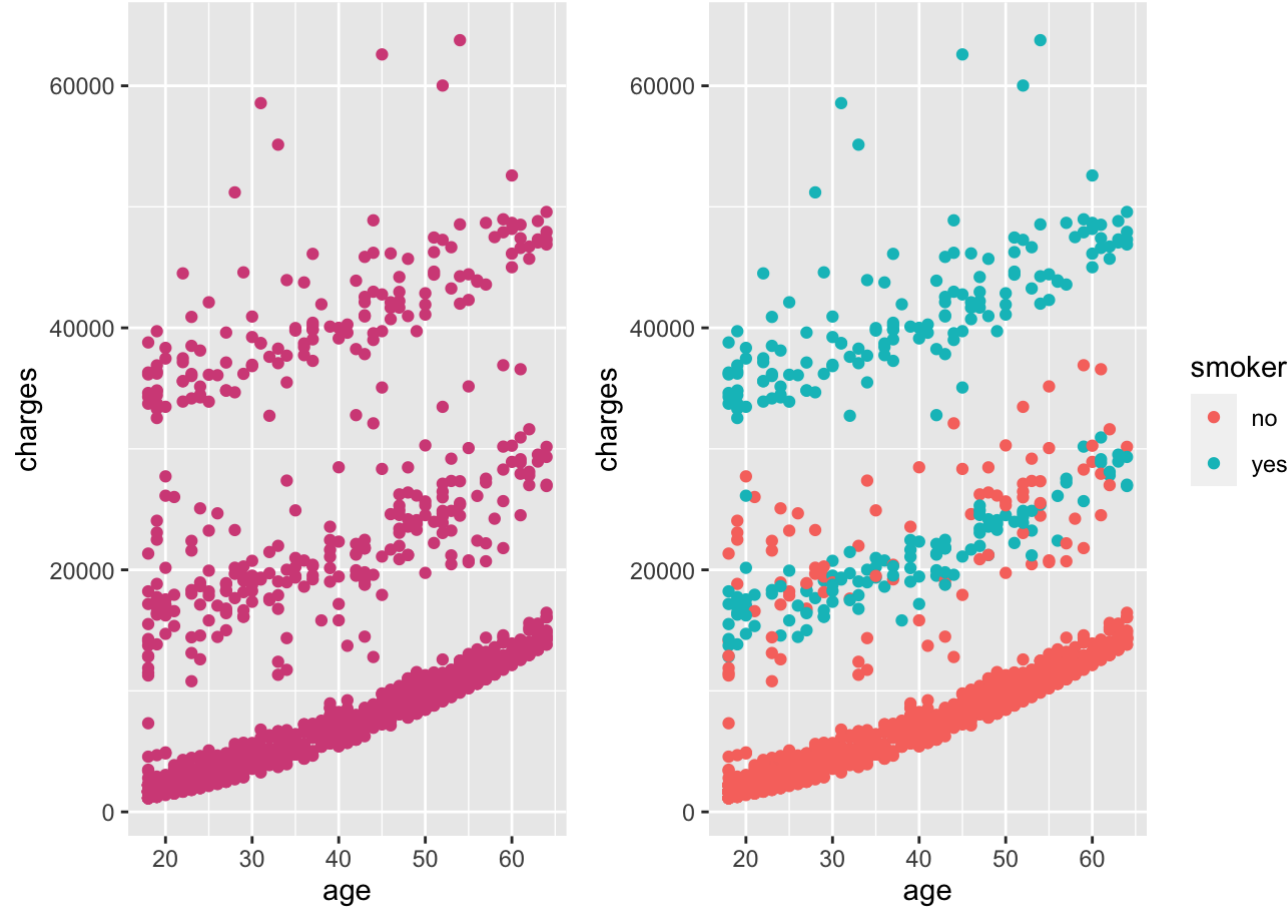
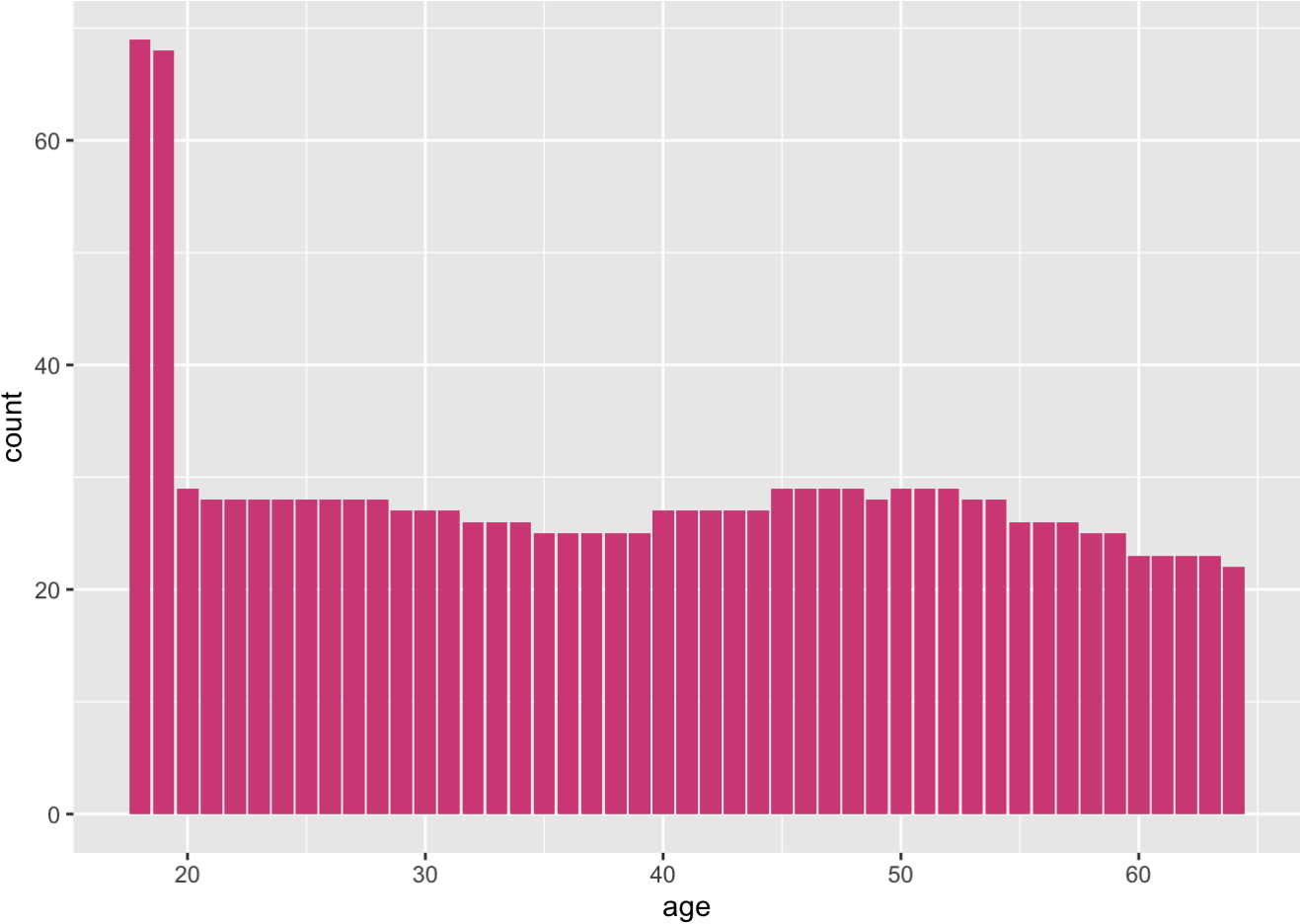
3.1insurance分布图



```
## # A tibble: 1 x 3
##   low_charges middle_charges high_charges
##   <int>         <int>         <int>
## 1      980         351           7
```

健康保险支出费用集中在1.5万美元以下，占比73.3%。1.5万–5万之间占比26.2%，5万以上占比0.5%。这三个区间段的支出费用差距较大，应采取不同的运营策略。

3.2年龄分布图

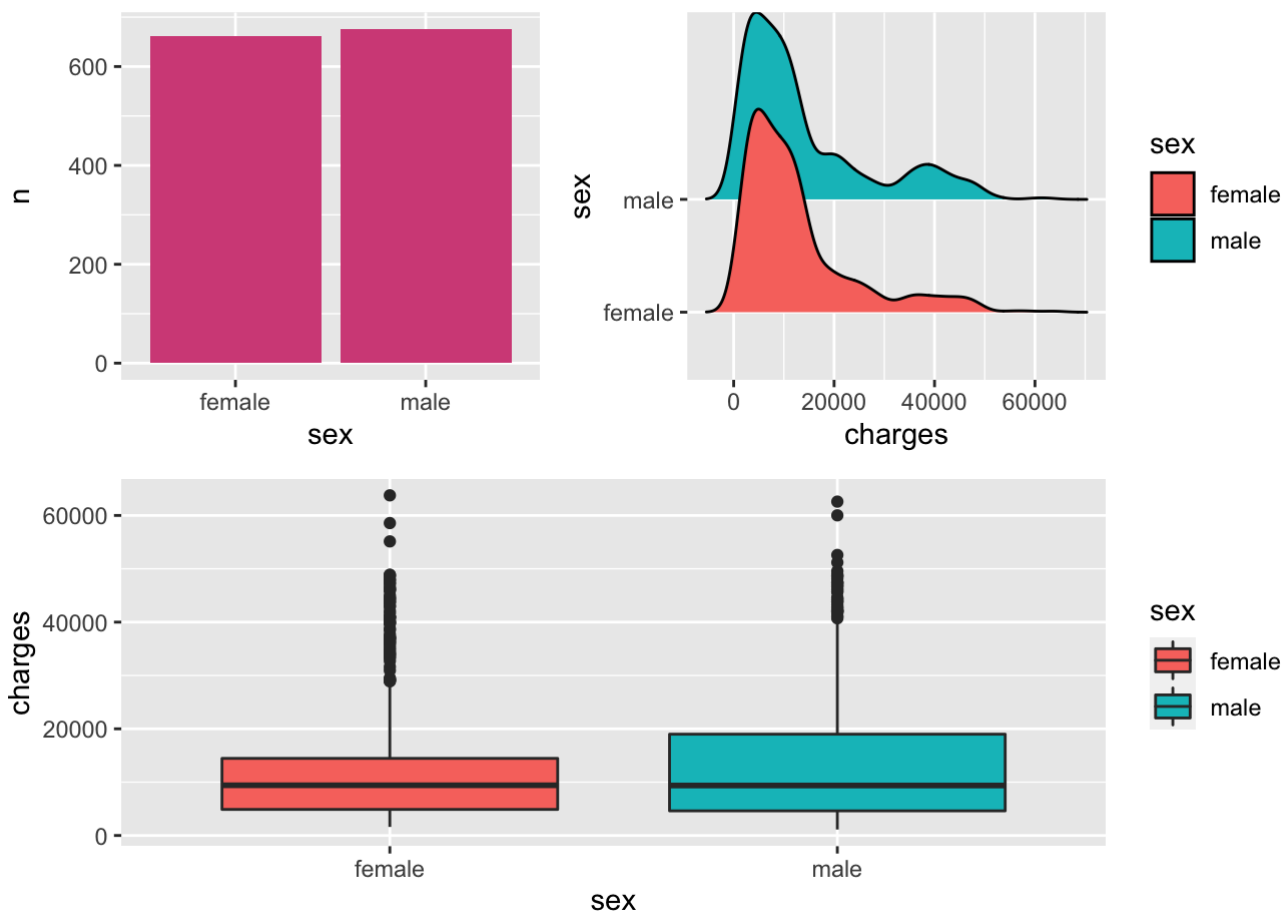


图形表明：

- 1. 年龄条形图显示，投保费用中20岁以下的人数占比最高，20岁以上人数分布均匀。
- 2. 抽烟群体的保险支出费用比非抽烟人群的保险支出费用要高，都成线性增长趋势。
- 3. 年龄和保险支出费用呈线性关系，但是这种关系应该收到其他因素的影响，表现为多层线性关系。
- 4. 依照是否吸烟对数据进行分组，支出费用小于20000美金的用户中，线性关系非常明显，数据分布集中，离群值较少。
- 5. 红色和绿色重叠部分，由于数据集中没有关于抽烟支数的记录，猜测可能是偶尔吸烟和不吸烟的人群。
- 6. 医疗费用支出最高的用户可能是老烟民。

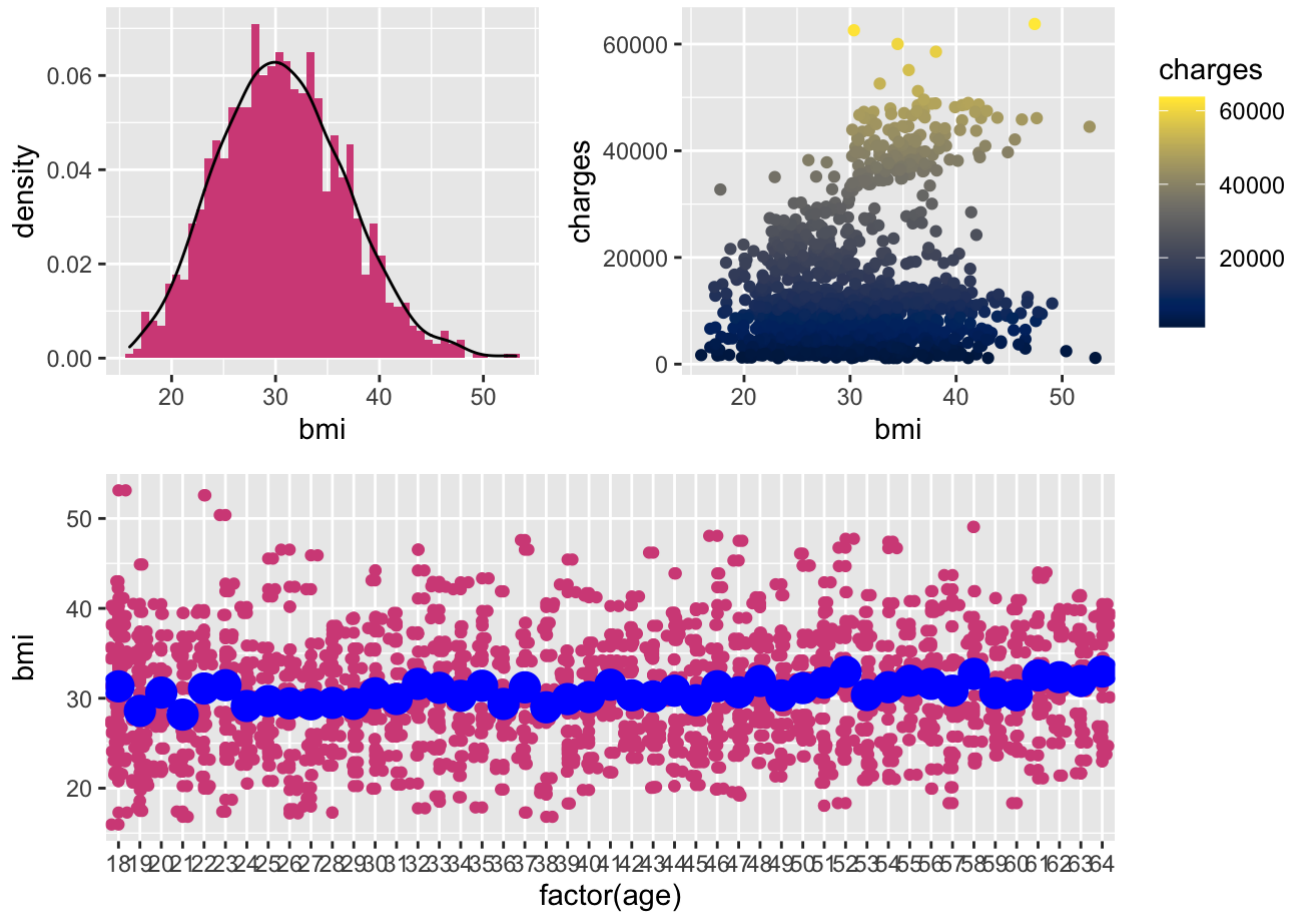
3.3性别分布图

Picking joint bandwidth of 2190



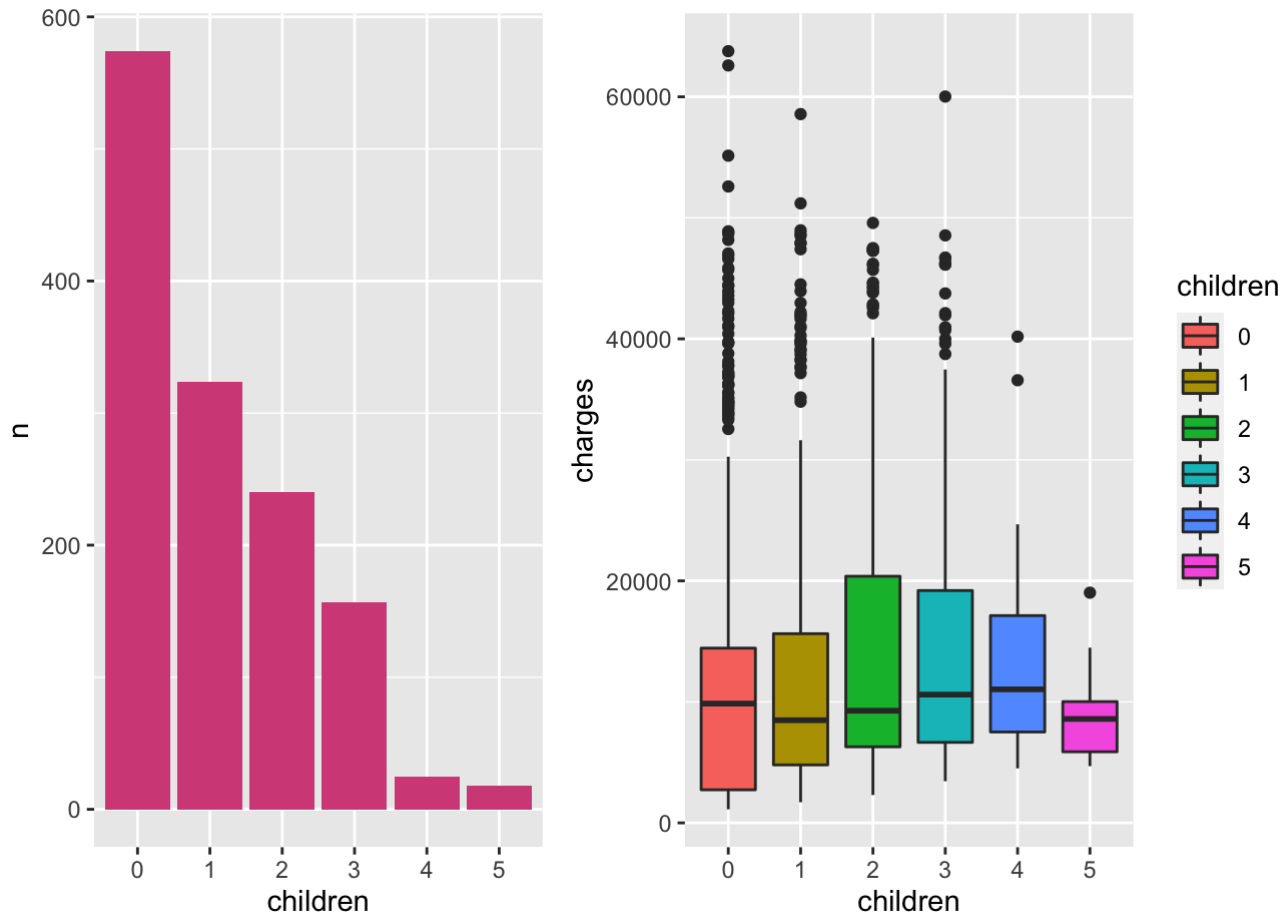
性别因素，样本中男女人数均衡，费用支出无大的差异。

3.4bmi分布图



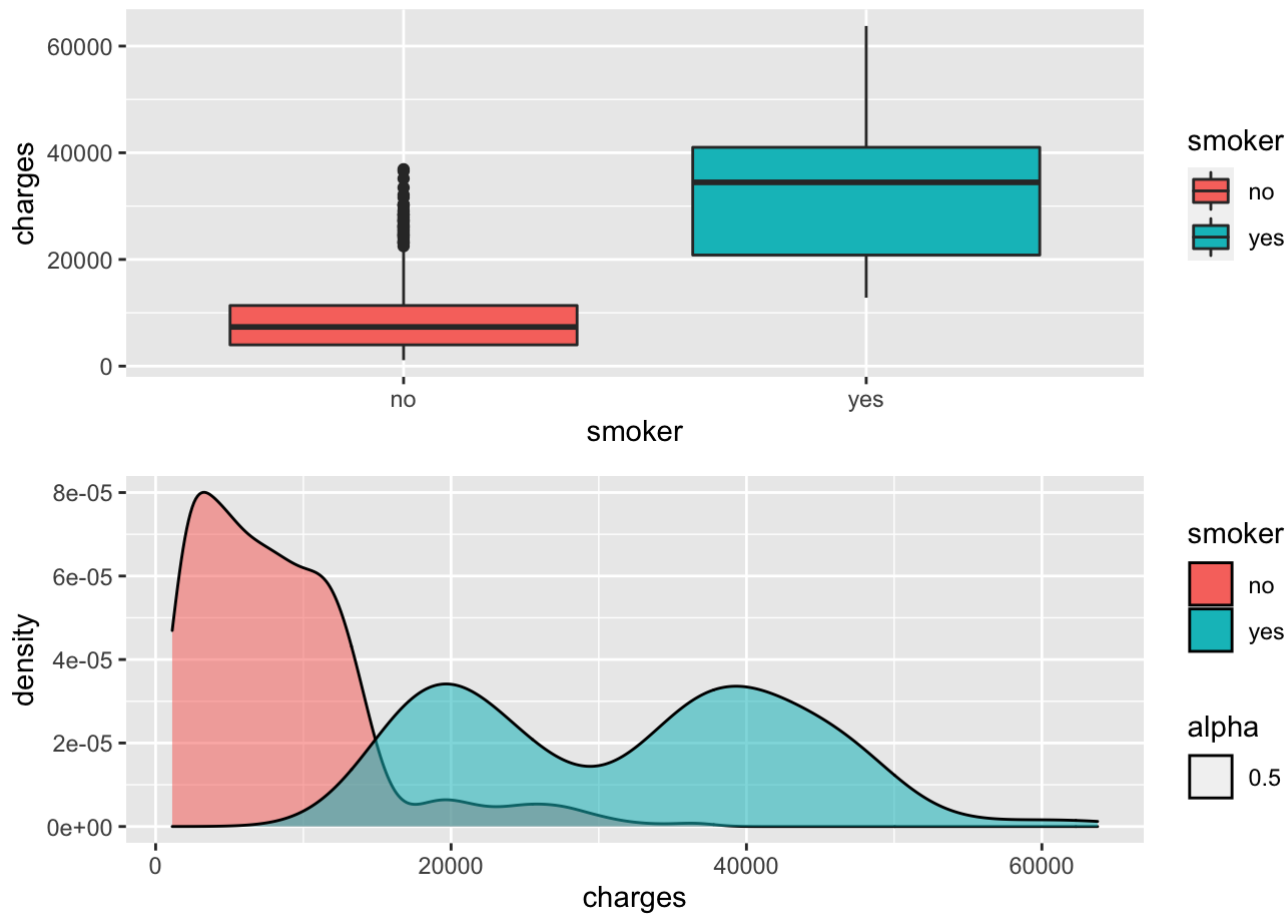
bmi数据呈常态分布，用户各个年龄段的平均bmi为30，bmi范围集中（25-35），bmi与保险支出看不到明显的线性关系。不能单纯的把bmi的高低作为保险费用高低的评价标准。
关于令人惊奇的是，bmi因素，并不像主观认定的，越肥胖的人，健康程度越差，健康费用支出越多。

3.5儿童个数分布图



没有儿童的用户大约占比44%。儿童数量越多，保险费用出现大金额的概率越低。

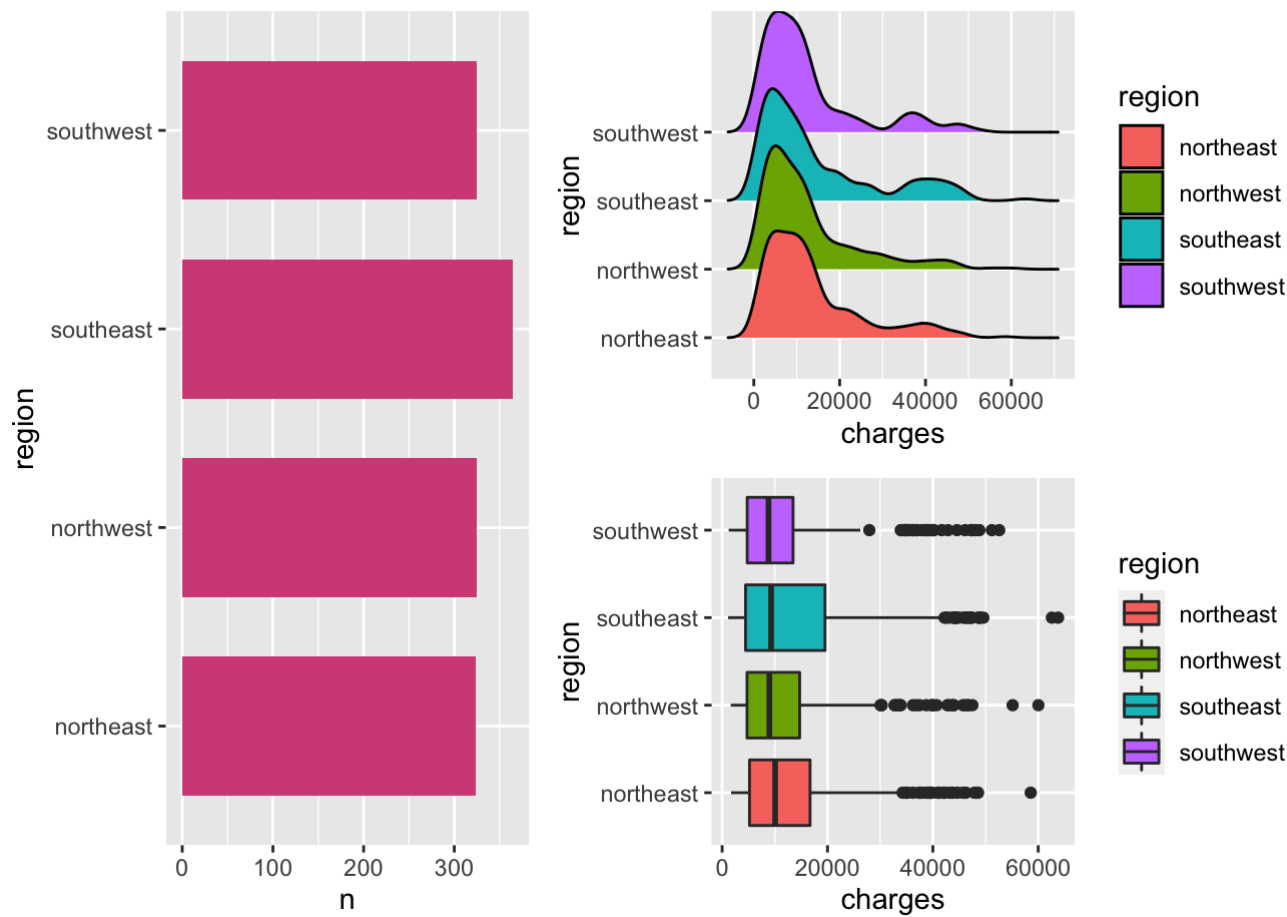
3.6吸烟因素分布图



费用支出在吸烟因素中的表现，区别非常大。非吸烟用户的支出费用在100-1600美元浮动；吸烟用户的波动范围为15000-55000。吸烟用户的保险支出费用最小值和非吸烟用户支出费用的最大值持平。可以清晰的看到制约健康保险费用的关键因素中，是否吸烟是其中重要因素之一。

3.7地区因素分布图

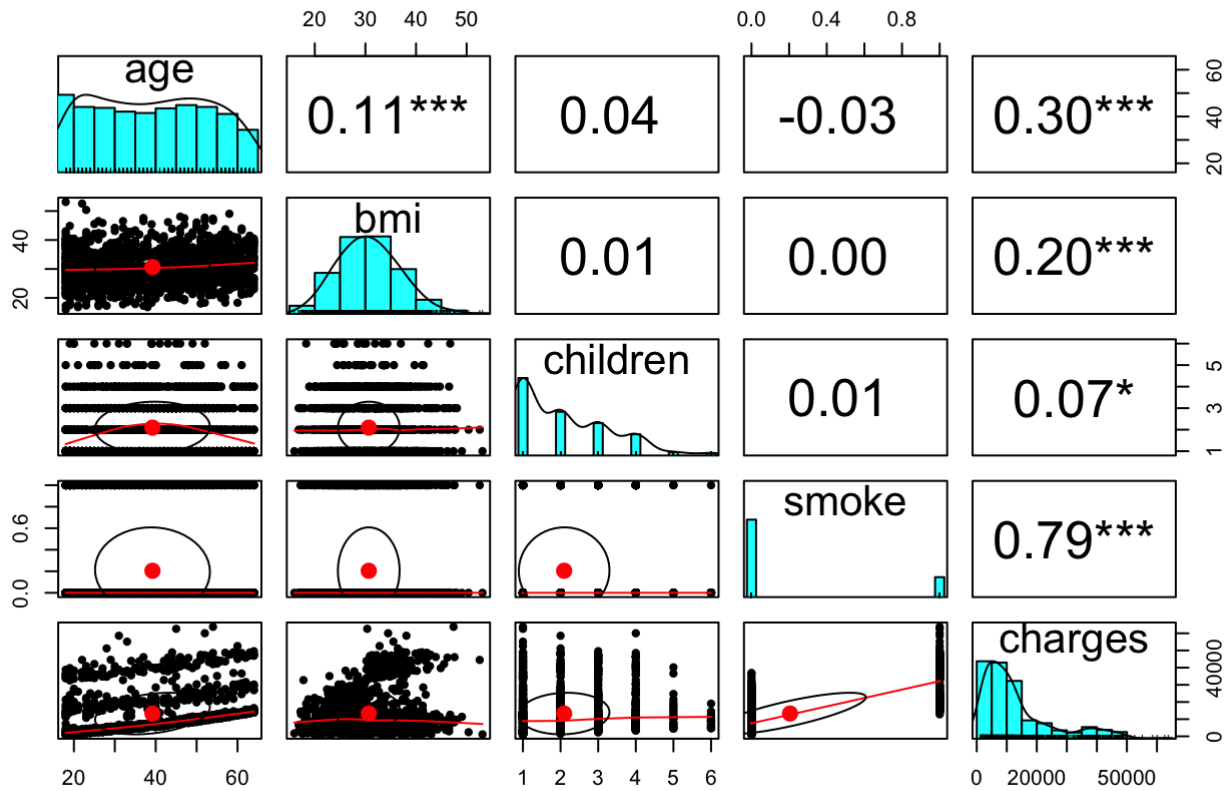
```
## Picking joint bandwidth of 2370
```



地区因素中，四个地区在费用支出上数据表现没有大的差异性。

四、相关性探索

Insurance Scatterplot Matrix



可以看到charges同smoke、age、bmi都呈现正相关，charges同smoke的相关系数达到了0.79。是否吸烟已经成了影响健康费用的决定性因素。

五、建立模型

5.1 多元线性模型

```
##
## Call:
## lm(formula = charges ~ smoker + age + bmi + children + region,
##     data = insu_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11620.3  -2883.5   -945.6   1513.0  29986.9
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -11977.26     984.79  -12.162 < 2e-16 ***
## smokeryes      23824.24     412.80   57.714 < 2e-16 ***
## age           257.30       11.91   21.609 < 2e-16 ***
## bmi           336.39       28.57   11.774 < 2e-16 ***
## children1      388.71     421.17    0.923 0.356211
## children2     1635.23     466.52    3.505 0.000471 ***
## children3      962.98     547.91    1.758 0.079055 .
## children4     2938.65    1238.56    2.373 0.017804 *
## children5     1106.45    1455.33    0.760 0.447227
## regionnorthwest -379.44     476.40   -0.796 0.425908
## regionsoutheast -1032.43     478.98   -2.155 0.031304 *
## regionsouthwest -952.16     478.00   -1.992 0.046577 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6057 on 1326 degrees of freedom
## Multiple R-squared:  0.7519, Adjusted R-squared:  0.7498
## F-statistic: 365.3 on 11 and 1326 DF,  p-value: < 2.2e-16
```

在 $p < 0.001$ 水平下，截距、吸烟人群、年龄、bmi和2个儿童的回归系数都非常显著。该模型可以解释75%的变异。

```
##
## Call:
## lm(formula = charges ~ smoker + age + bmi, data = insu_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12415.4  -2970.9   -980.5   1480.0  28971.8
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11676.83     937.57  -12.45  <2e-16 ***
## smokeryes    23823.68     412.87   57.70  <2e-16 ***
## age          259.55       11.93   21.75  <2e-16 ***
## bmi          322.62       27.49   11.74  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6092 on 1334 degrees of freedom
## Multiple R-squared:  0.7475, Adjusted R-squared:  0.7469
## F-statistic: 1316 on 3 and 1334 DF, p-value: < 2.2e-16
```

模型2的决定系数为74.8，各变量的回归系数都非常显著。

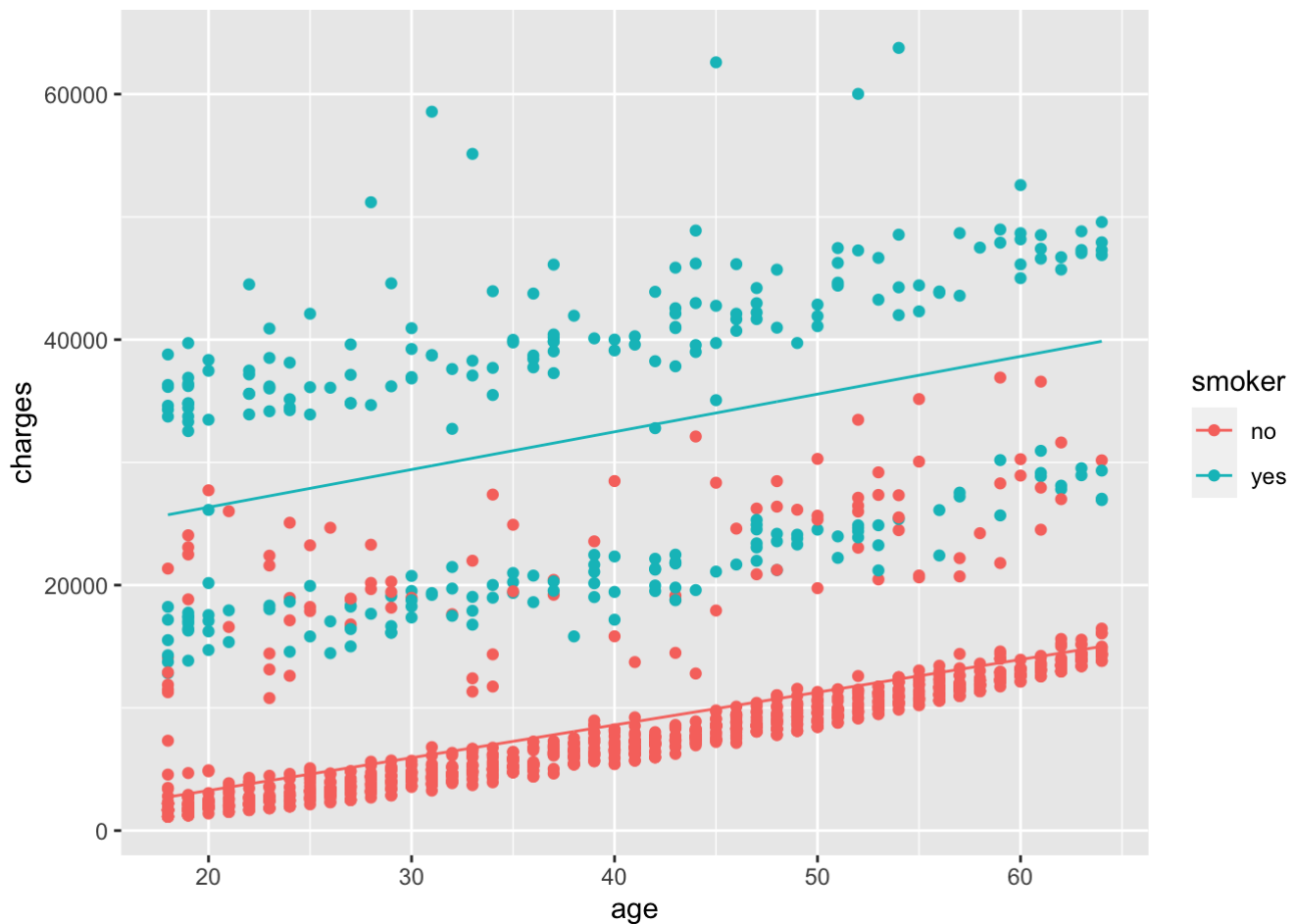
5.2多重线性模型

```
## boundary (singular) fit: see ?isSingular
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: charges ~ age + (1 + age | smoker)
## Data: insu_df
## REML criterion at convergence: 27234.39
## Random effects:
## Groups Name Std.Dev. Corr
## smoker (Intercept) 9963.72
## age 18.18 1.00
## Residual 6396.79
## Number of obs: 1338, groups: smoker, 2
## Fixed Effects:
## (Intercept) age
## 9066 287
## optimizer (nloptwrap) convergence code: 0 (OK) ; 0 optimizer warnings; 1 lme4 warning
s
```

以是否吸烟对把数据分为两组，可以看到是否吸烟对于年龄因素在保险费用支出上的变化效应。

5.2.1以smoker为分组对age变量的多重线性模型可视化



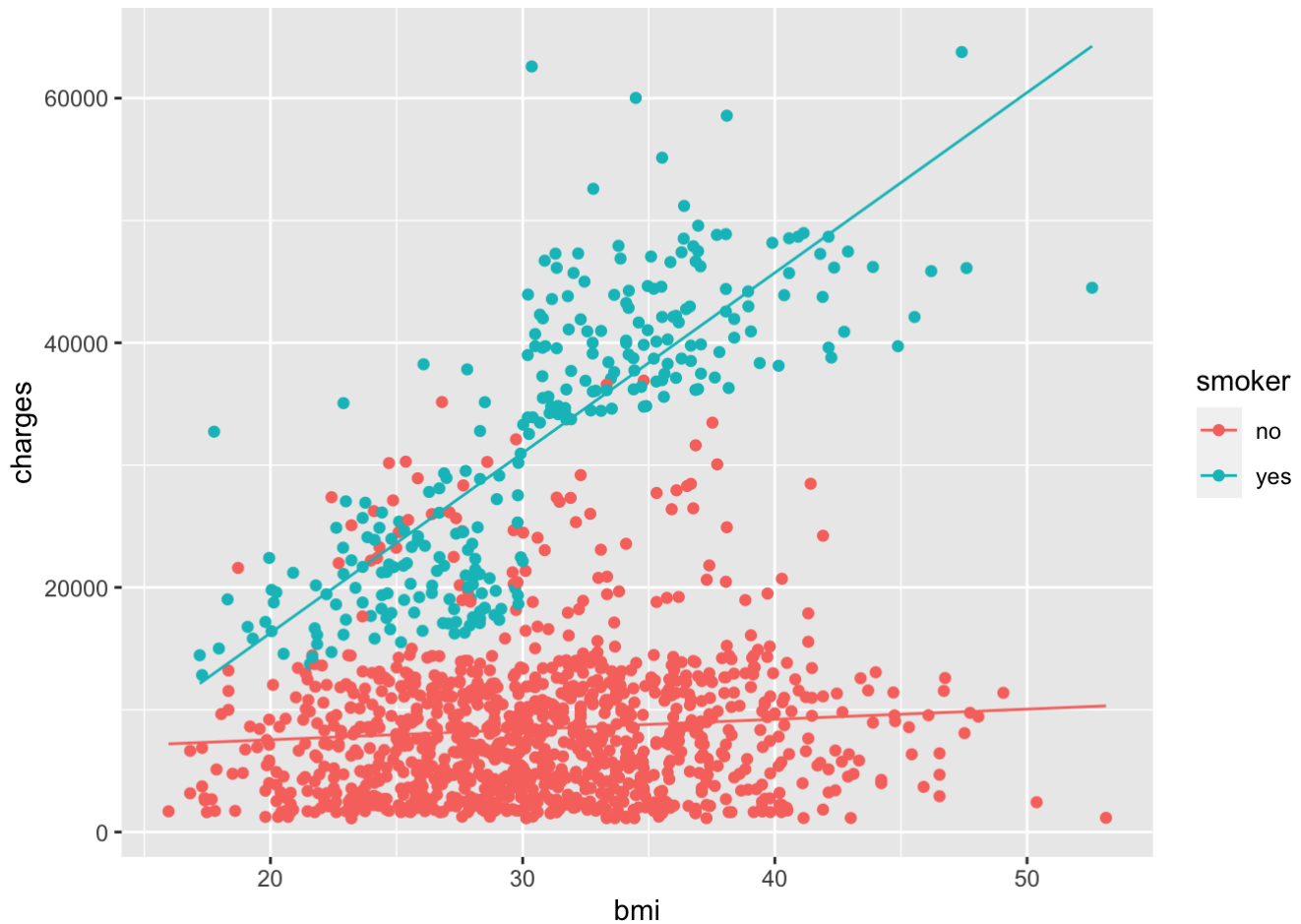
该模型对不抽烟的用户拟合的还是比较好的，但是对于吸烟的用户拟合直线明显感觉有些粗旷。或许与吸烟的量以及其他因素有关，导致该模型拟合的绿色直线，不够好。

```
## boundary (singular) fit: see ?isSingular
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: charges ~ bmi + (1 + bmi | smoker)
## Data: insu_df
## REML criterion at convergence: 27132.37
## Random effects:
## Groups Name Std.Dev. Corr
## smoker (Intercept) 13479.9
## bmi 982.6 -1.00
## Residual 6158.5
## Number of obs: 1338, groups: smoker, 2
## Fixed Effects:
## (Intercept) bmi
## -3652.3 778.1
## optimizer (nloptwrap) convergence code: 0 (OK) ; 0 optimizer warnings; 1 lme4 warning
s
```

该模型以不吸烟组作为参数对照，可以看到是否吸烟对于bmi因素在保险费用支出上的变化效应。

5.2.2以smoker为分组对bmi变量的多重线性模型可视化



这个模型的解释能力很好，吸烟人群的随着bmi的增加，支付费用高速增加；非吸烟人群随着bmi的增加，支出增加的相对要缓慢些。

结论

1. 美国居民医疗费用支出集中15000美元以下，占比73.3%。1.5万-5万之间占比26.2%。这两部分共占据了99.5%的医疗费用支出情况。
2. 健康保险费用在性别和地区上没有明显的差异。
3. 从模型上看，医疗费用支出同bmi、smoker、age这三个变量表现出较高的相关性。其中，吸烟与否对与费用影响很大。具体来说吸烟人群中年龄大又肥胖的人群，费用支出非常高。对于非吸烟人群，哪怕胖一些，支出费用也不会非常高。
4. 在制定健康保险策略是，应从总体上把用户分为吸烟或者非吸烟用户，然后再去看用户的年龄和bmi情况，这样可以更优的使用保险资源。同时，应鼓励人们戒烟，并适当的进行减肥活动。