统计机器学习课后作业 3

陈劭涵 17300180049

2020 年 10 月 25 日

1 问题 1

数据导入:

```
data = read.csv("D:/大数据学院文件资料/2020秋课程/机器学习/pj1/data.csv")
```

汇总统计:

```
summary(data)
```

统计结果:

```
      rent          bedroom        livingroom      bathroom
Min   :1150   Min   :2.000   Min   :1.00   Min   :1.000
1st Qu.:2240   1st Qu.:2.000   1st Qu.:1.00   1st Qu.:1.000
Median :2690   Median :3.000   Median :1.00   Median :1.000
Mean   :2798   Mean   :2.996   Mean   :1.01   Mean   :1.027
3rd QU.:3230   3rd Qu.:4.000   3rd Qu.:1.00   3rd Qu.:1.000
Max.   :6460   Max.   :5.000   Max.   :2.00   Max.   :2.000
      area          room        floor_grp      subway
Min   : 5.00   次卧:2860   低楼层:1679   否:815
1st Qu.:10.00   主卧:2289   高楼层:1592   是:4334
Median :12.00               中楼层:1878
Mean   :12.85
3rd Qu.:15.00
Max.   :30.00
    region          heating
朝阳   :1317   集中供暖:4197
通州   :819    自采暖  :952
昌平   :702
丰台   :581
```

```
海淀      :424
大兴      :361
(other):945
```

统计结果解读：

1、月租金最高达 6460 元，最低仅 1150 元，平均 2798 元，中位数 2690 元；

2、卧室数量最低 2 个，最高 5 个，平均达到 3 个；

3、厅数最低 1 个，最高 2 个，平均 1 个，说明绝大多数的租房只有一个厅，极少数有两个厅；

4、卫生间数最低 1 个，最高 2 个，平均 1 个，说明绝大多数的租房只有一个卫生间，极少数有两个卫生间；

5、租房面积最低 5 平方，最高 30 平方，平均 12 平方，第三四分位数 15 平方，说明租房面积普遍比较小，大部分租房面积不超过 15 平方;

6、租赁房间类型，次卧 2860 个样本，主卧 2289 个样本。说明租房类型中次卧略大于主卧；

7、楼层分布上，低楼层 1679 个，中楼层 1878 个，高楼层 1592 个；楼层分布总体上较为均匀，以中楼层略微居多;

8、是否临近地铁方面，4334 个租房临近地铁，815 个租房不临近地铁，说明大部分的租房都靠近地铁线路;

9、城区分布上，朝阳区占最大比例，有 1317 例，超过其他城区的租房数量;

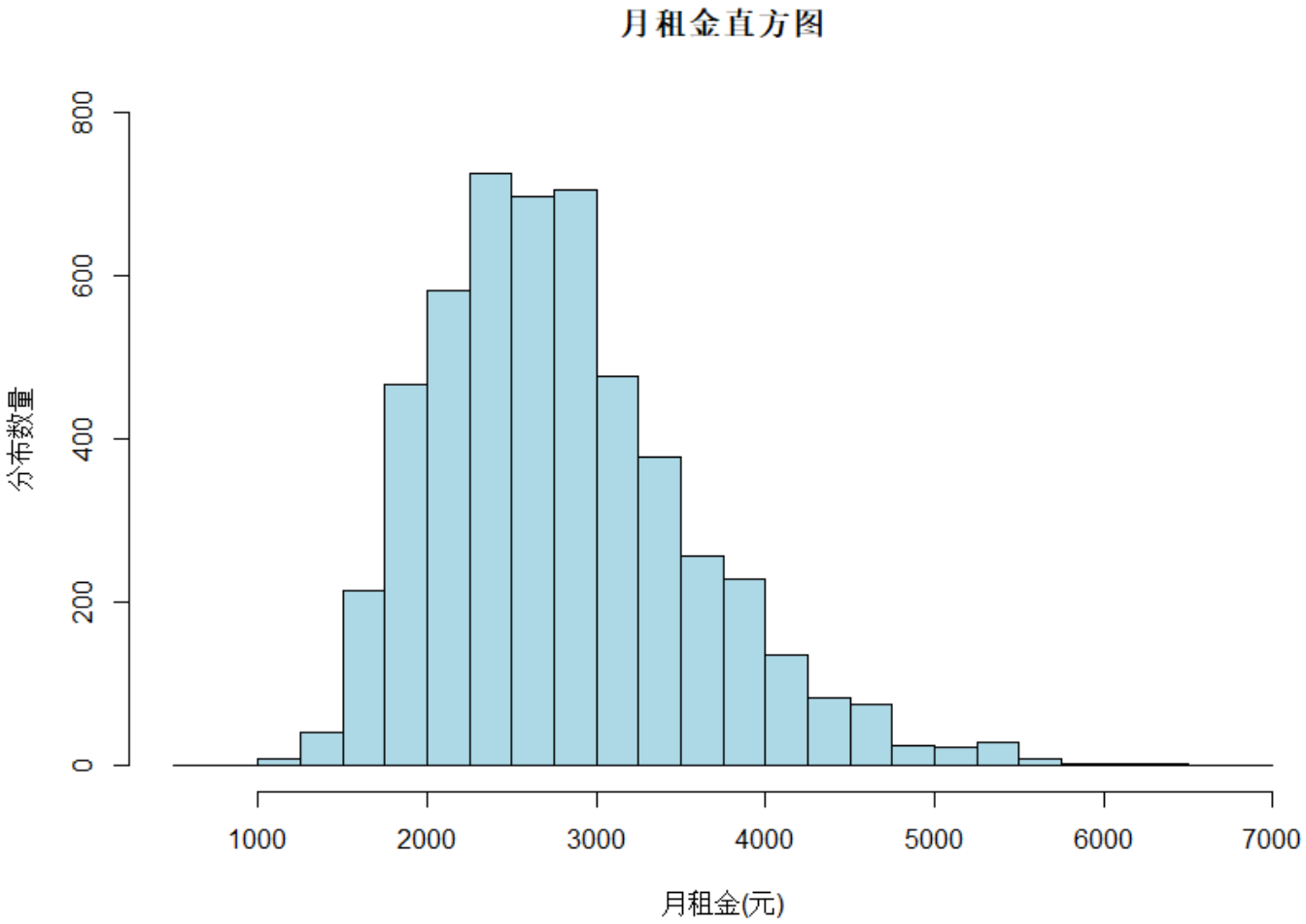
10、供暖情况上，有 4197 例租房采用集中供暖的方式，仅 952 例租房采用自采暖的方式。说明大部分的租房都采用集中供暖的形式。

2 问题 2

绘制直方图:

```
# 绘制直方图
hist(data$rent,seq(from=500,to=7000,by=250),col="lightblue",main="月租金直方图",
xlab = "月租金(元)",ylab="分布数量",ylim=c(0,700))
```

直方图：



解读:

1、根据第一题的数据统计，月租金最低 1150 元/月，最高 6460 元/月，平均 2798 元/月，中位数 2690 元/月；

2、月租金分布是右偏的，大部分租金都集中分布在 1500/月 ~4000 元/月之间。分布水平的峰值出现在月租金 2200 至 3000 元范围内, 包含近 2200 个样本点，超过其他区间范围内的样本量；

3 问题 3

计算不同地区平均租金:

```
tapply(data$rent,data$region,mean)
```

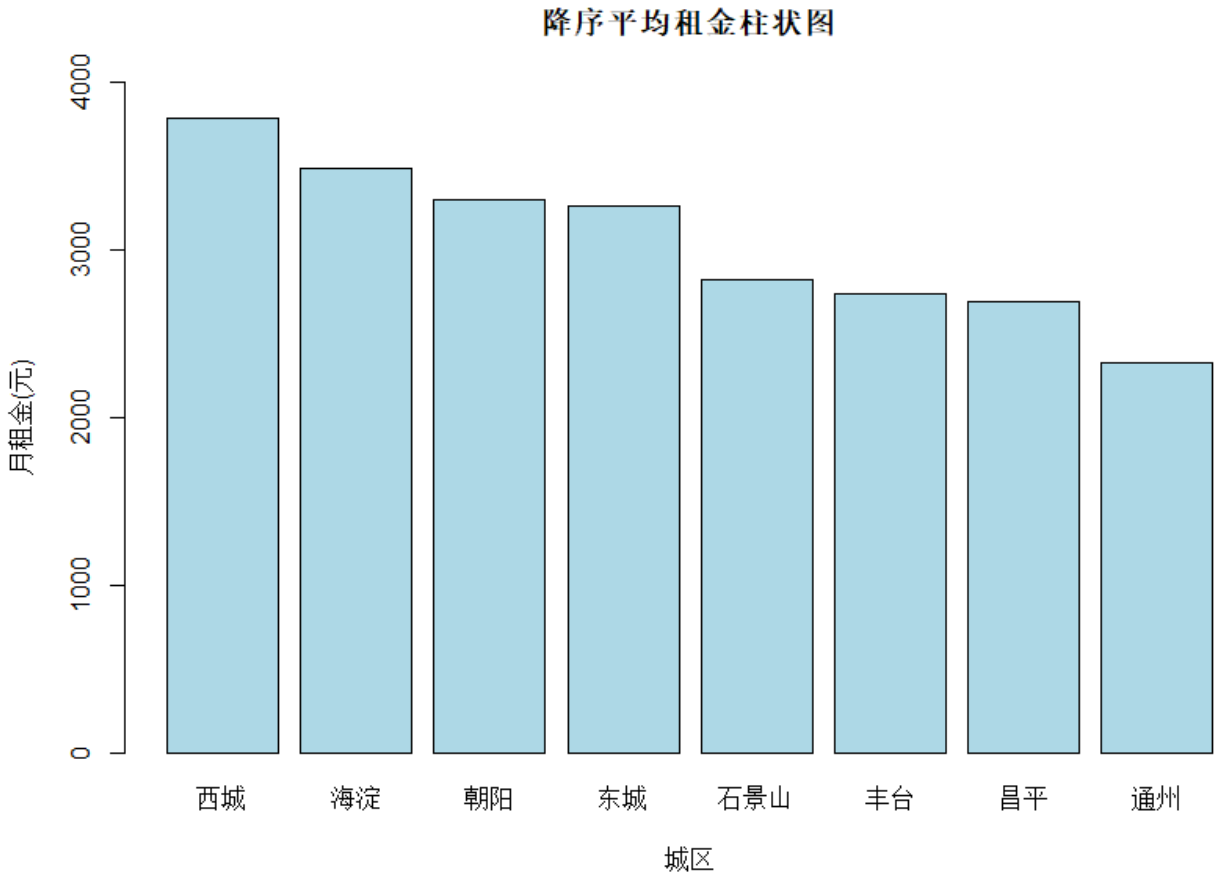
不同地区平均租金如下：

```
昌平      朝阳      大兴      东城      房山      丰台      海淀      石景山    顺义
2693.376 3302.103 2241.136 3262.872 1751.257 2734.286 3490.920 2819.845 2111.497
通州      西城
2327.216 3784.792
```

选取平均租金最高的 8 个城区，绘制降序平均租金柱状图:

```
# 绘制降序柱状图
t=tapply(data$rent,data$region,mean)
t= sort(t, decreasing = TRUE)[1:8]
bar=barplot(t,col="lightblue",main="降序平均租金柱状图",
xlab="城区",ylab="月租金(元)",ylim=c(0,4000))
```

结果如下:



简要分析:

1、总体上，在图中所示部分，平均租金最高不超过 4000 元/月，最低不小于 2000 元/月，平均租金的平均水平大约在 2700 至 3000 元/月之间; 大兴、房山、顺义三个城区由于租金太低，排在第八位之后，不出现在降序柱状图中

2、在平均租金最高的 8 个城区中，平均租金最高的城区是西城区，达到 3784.792 元/月，平均租金最低的城区是通州区，为 2327.216 元/月。而如果考虑所有计算出的平均租金，有三个城区比通州区还要低，最低的是房山区，为 1751.257 元/月；

3、不同城区的平均租金从高到低的变化总体上较为渐进。但在西城区到海淀区，以及东城区到石景山区之间出现较为明显的月租金变化，从昌平区到通州区也有明显的租金下降。而朝阳区和东城区的月租金比较接近，石景山、丰台、昌平区的月租金比较接近；

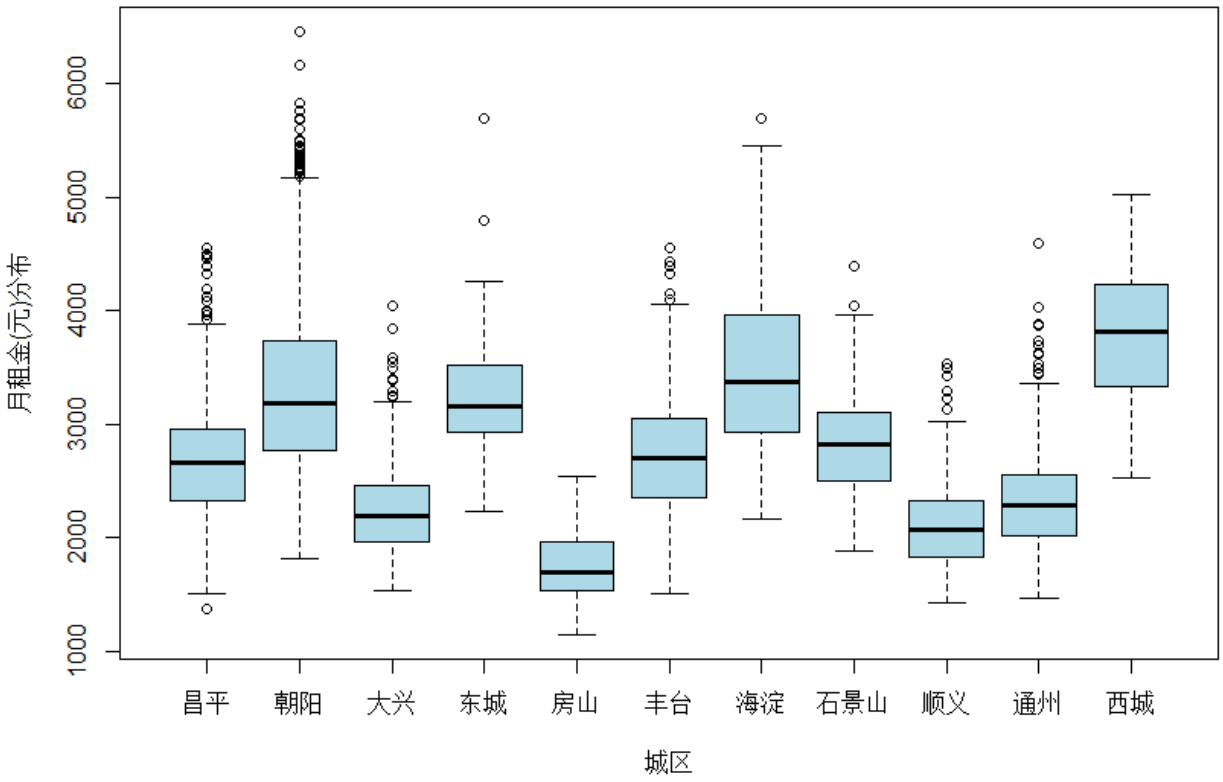
4、从降序柱状图上看，月租金总体上可以分为四个水平，分为对应西城，海淀 + 朝阳 + 东城，石景山 + 丰台 + 昌平，通州这四部分城区。

4 问题 4

绘制月租金（rent）-城区（region）分组箱线图:

```
# 绘制箱线图
boxplot(data$rent~ data$region,col="lightblue",horizontal = F,
xlab="城区",ylab="月租金(元)分布",plot = T)
```

结果如下:



简要解读：

1、箱线图显示的月租金水平分布与上一题所展示的平均租金水平分布有较为相近之处。可以看到，西城区从月租金分布上是总体水平最高的城区，房山区是最低的；西城区，丰台区和东城区等城区的月租金分布都较为均匀；

2、从上边缘上看，海淀区和朝阳区的月租金上边缘是最高的，甚至超过平均水平最高的西城区。这两个城区同时也是月租金分布范围最广 (上下边缘差距最大) 最分散的城区，月租金分布也最不均匀。

这可能和城区和城区内部特殊的区位因素有关。其他城区的月租金分布相对集中一些。房山区的月租金水平在上边缘和下边缘都是最低的；

3、西城区，房山区，东城区和石景山区等城区的离群值较少，分布也较为稳定。朝阳区，长兴区，大兴区和通州区等城区的离群值较多，并且这些离群值大部分都是高于上边缘的，说明在这些城区的某些位置可能由于各种区位或政策因素，导致这些位置的房价远远高于正常水平；

5 问题 5

建立以月租金 (rent) 为因变量，其余为自变量的线性回归模型，并提前指定基准组:

```
# 设立基准组并转化为factor变量
data$room=factor(data$room,levels=c("次卧","主卧"))
data$floor_grp=factor(data$floor_grp,levels=c("低楼层","中楼层","高楼层"))
data$subway=factor(data$subway,levels=c("否","是"))
data$region=factor(data$region,levels=
c("石景山","昌平","朝阳","大兴","东城","房山","丰台","海淀","顺义","通州","西城"))
data$heating=factor(data$heating,levels=c("自采暖","集中供暖"))
# 线性回归与结果输出
reg=lm(formula=rent~.,data=data)
summary(reg)
```

线性回归模型结果如下:

```
Call:
lm(formula = rent ~ ., data = data)

Residuals:
Min        1Q    Median        3Q       Max
-1617.48  -275.26   -23.21   248.40  2967.39

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)     1593.22405   79.93979  19.930  < 2e-16 ***
bedroom          -90.55187    8.73965 -10.361  < 2e-16 ***
livingroom      -168.01859   71.84304  -2.339 0.019390 *
bathroom         182.10825   43.63344   4.174 3.05e-05 ***
```

```
area                 76.69198    1.96317   39.065  < 2e-16 ***
room主卧              0.08443   16.57165    0.005 0.995935
floor_grp中楼层     -55.59610   15.29704   -3.634 0.000281 ***
floor_grp高楼层     -24.98532   15.95019   -1.566 0.117303
subway是            280.44440   17.73655   15.812  < 2e-16 ***
region昌平           57.19973   34.07111    1.679 0.093245 .
region朝阳          631.69500   31.69727   19.929  < 2e-16 ***
region大兴         -421.98157   37.42529  -11.275  < 2e-16 ***
region东城          565.01670   55.05629   10.263  < 2e-16 ***
region房山         -811.82173   45.02264  -18.031  < 2e-16 ***
region丰台          117.59678   34.57808    3.401 0.000677 ***
region海淀          878.86347   36.51087   24.071  < 2e-16 ***
region顺义         -450.95937   39.11271  -11.530  < 2e-16 ***
region通州         -373.00568   32.65046  -11.424  < 2e-16 ***
region西城          938.86226   54.33181   17.280  < 2e-16 ***
heating集中供暖     155.79086   17.13452    9.092  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 454.1 on 5129 degrees of freedom
Multiple R-squared:  0.6466,    Adjusted R-squared:  0.6453
F-statistic: 493.9 on 19 and 5129 DF,  p-value: < 2.2e-16
```

观察最终的回归系数并对系数进行解释:

1、截距项为 1593.22，随着自变量每增加一单位，可以看到：
卧室平均每增加一间，月租金减少 91 元;
厅数平均每上升一间，月租金减少 168 元，但显著性水平一般;
卫生间平均每增加一个，月租金增加 182 元;
面积平均每增加一平米，月租金增加 77 元;
主卧对月租金的影响比较小，显著性水平很低，主卧比次卧的月租金略微低一点;
中楼层比低楼层的月租金平均少 56 元，高楼层比低楼层的月租金平均少 25 元;
有地铁的房子月租金平均比没有地铁的要高 280 元;
昌平，朝阳，东城，丰台，海淀，西城城区的平均月租金要比石景山区高 57,631,565,118,879,939 元;
大兴，房山，顺义，通州城区的平均月租金要比石景山区低-422，-812，-451，-373 元; 不难想石景山区的租金位于一个中间的位置，这也与前面一题的结果相符
集中供暖的房子月租金平均比自供暖的房子高 155.8 元;

2、根据回归模型的系数，可以看到，除了 livingroom，room 主卧，floor-grp 高楼层，以及 region 昌平这几个自变量，其他的自变量的显著性水平都非常高 (0.001)，说明他们对于模型与因变量具有较强解释作用；在显著性水平较弱的这几个变量中，livingroom 的显著性还是比较强的 (0.05)，但剩余的三个自变量的显著性就比较弱了，说明他们对于因变量的解释性比较弱；

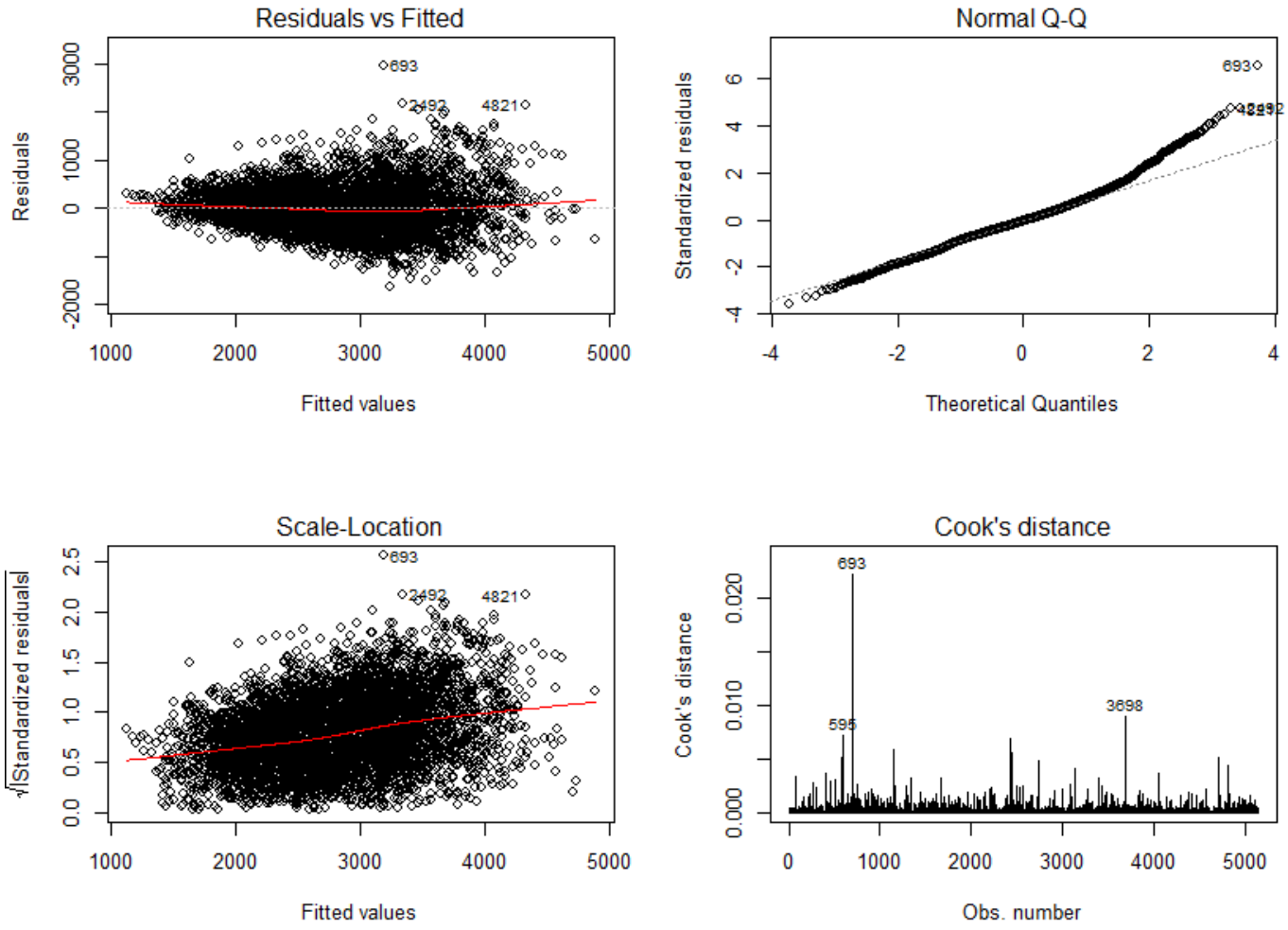
3、模型的 R-squared 与 Adjusted-R-squared 的值均在 0.64 左右，表示的是模型的拟合程度。从结果上看，线性回归模型对观测值的拟合程度一般，不是很高；

4、模型的 F 检验的 p-value 很小，所以可以认为方程在 P=0.001 的水平上通过显著性检验；

我们再简单看看回归诊断的情况：

```
# 绘制回归诊断图
par(mfrow=c(2,2))
plot(reg,which=c(1:4))
```

结果:



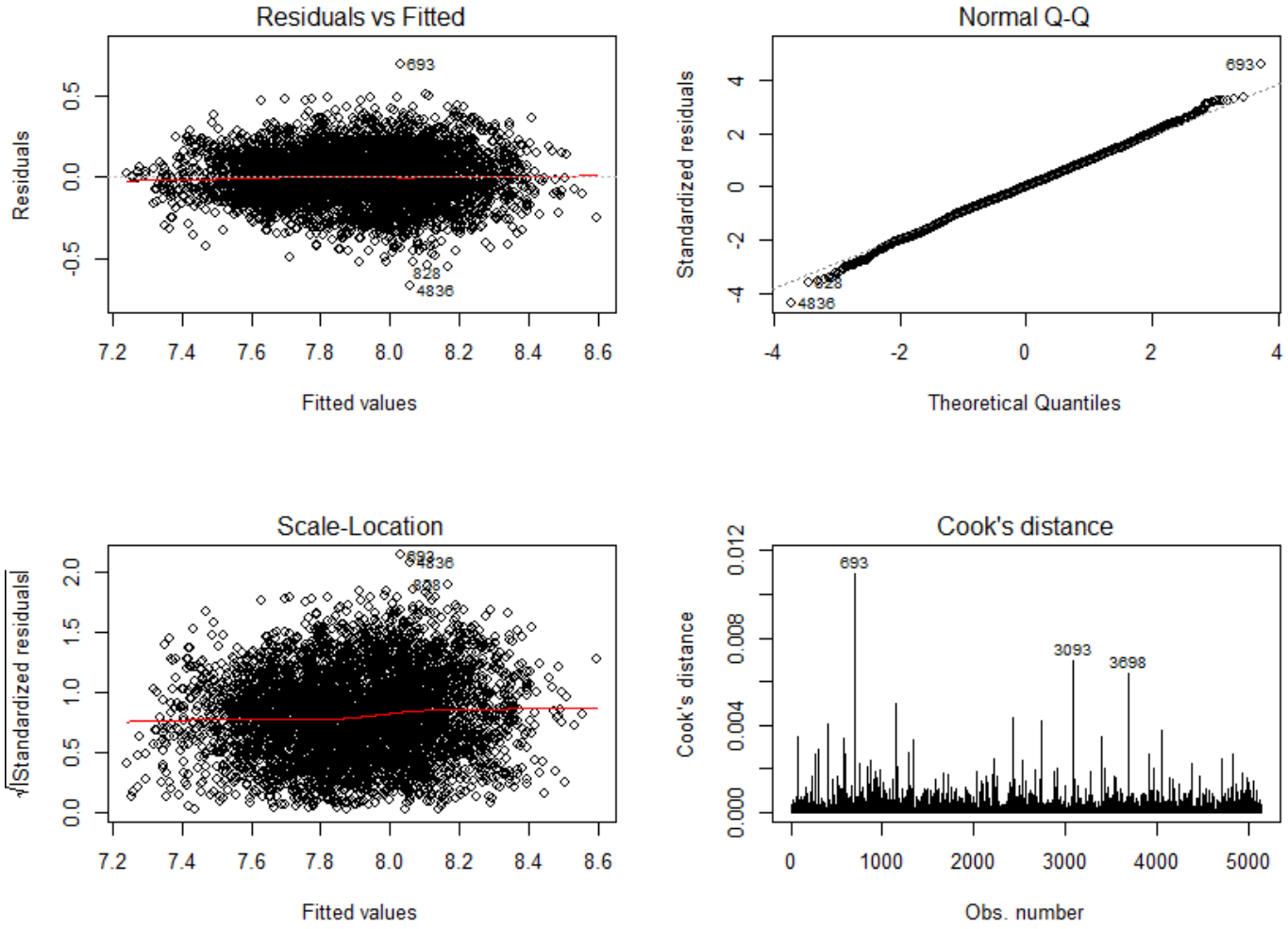
根据图表我们可以看到：

1、模型的残差与拟合结果显示，模型 0 总体上保持线性，但是存在异方差现象，可以考虑变换因变量；

2、模型的正态性诊断结果显示，模型的正态性一般，在尾部发生较大偏离；

3、模型的库克距离诊断显示，模型中有一些点对模型的影响较大，这些点可能是异常点

如果我们将因变量取对数，重新进行回归诊断，结果如下：



可以看到取对数确实减少了异方差现象，并且使得正态诊断结果更加良好

6 问题 6

对上面的回归模型利用 BIC 准则进行变量选择并解读结果：

```
reg1=step(lm(rent~.,data=data),direction="both",trace=1,keep=NULL,
steps=2000,k=log(nrow(data)))
summary(reg1)
```

变量选择后的结果:(BIC 准则下取参数 k=log(n))

```
Call:
lm(formula = rent ~ bedroom + bathroom + area + subway + region +
heating, data = data)
```

```
Residuals:
Min        1Q    Median        3Q       Max
-1611.79  -271.91   -25.92   248.37  2980.50

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)      1440.449     59.027  24.403  < 2e-16 ***
bedroom           -91.457      8.547 -10.701  < 2e-16 ***
bathroom          141.751     40.550   3.496 0.000477 ***
area               76.527      1.533  49.930  < 2e-16 ***
subway是          281.206     17.743  15.849  < 2e-16 ***
region昌平         58.700     34.104   1.721 0.085276 .
region朝阳        634.860     31.729  20.009  < 2e-16 ***
region大兴       -418.866     37.422 -11.193  < 2e-16 ***
region东城        565.767     55.124  10.264  < 2e-16 ***
region房山       -808.478     45.011 -17.962  < 2e-16 ***
region丰台        119.461     34.606   3.452 0.000561 ***
region海淀        880.960     36.546  24.106  < 2e-16 ***
region顺义       -456.412     39.038 -11.691  < 2e-16 ***
region通州       -372.159     32.685 -11.386  < 2e-16 ***
region西城        935.072     54.387  17.193  < 2e-16 ***
heating集中供暖   154.970     17.150   9.036  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 454.7 on 5133 degrees of freedom
Multiple R-squared:  0.6453,    Adjusted R-squared:  0.6443
F-statistic: 622.7 on 15 and 5133 DF,  p-value: < 2.2e-16
```

解读:

1、从变量选择后的 formula 中我们看到，留下来的变量是 bedroom, bathroom, area, subwa, region, heating。剩下的变量 livingroom, room 主卧和 floor-grp 等被变量选择过程所筛除；

2、从变量选择后的显著性系数中可以看到，除了 region 昌平之外的所有变量都达到了非常高的显著性水平 (0.001)，region 昌平也有 0.1 的显著性水平。原本存在的 livingroom, room 主卧和 floor-grp 等变量在原线性回归中显著性水平比较低，被筛除；

3、变量选择后，R-squared 和 Adjusted-R-squared 并没有提升，依然在 0.64 左右；这是可以理解的因为 R-squared 会随解释变量的增多而上升，但变量选择减少了解释变量的个数;

4、F 检验表示方程的显著性水平较高;

7 问题 7

对 BIC 选择后的模型进行五折交叉验证 (是没有取对数的原本模型):

```
# k折交叉验证函数(自定义)
pred.cv<-function(dat,k)
{
        ind=sample(1:k,nrow(dat),replace=T)
        pred_cv=rep(0,nrow(dat))
        for (i in 1:k)
        {
                ii=which(ind==i)
                obj=lm(rent ~ bedroom + bathroom + area + subway + region +
                heating,data=dat[-ii,])
                pred_cv[ii]=predict(obj,data[ii,])
        }
        rmse=sqrt(mean((pred_cv-data$rent)^2))
        return(list(pred_cv=pred_cv,rmse=rmse))
}

# 计算RMSE
set.seed(123)
rmses=rep(0,50)
for (i in 1:50)
{
        cat(i,"\r")
        pred_cv=pred.cv(dat=data,k=5)
        rmses[i]=pred_cv$rmse
}
mean(rmses)
```

验证结果 (rmse 值):

```
> mean(rmses)
[1] 455.5295
```

评估模型结果并进行解读:

从 RMSE 的值可以看到，RMSE 值较大，说明当前的线性模型的解释效果不是非常好，预测结果的具有较大的偏差，距离 0 也比较远; 于是我们尝试对租金取对数，进行对数线性模型预测:

```
# k折交叉验证函数(自定义)
pred.cv<-function(dat,k)
{
        ind=sample(1:k,nrow(dat),replace=T)
        pred_cv=rep(0,nrow(dat))
        for (i in 1:k)
        {
                ii=which(ind==i)
                obj=lm(log(rent) ~ bedroom + bathroom + area + subway + region +
                heating,data=dat[-ii,])
                pred_cv[ii]=predict(obj,data[ii,])
        }
        rmse=sqrt(mean((pred_cv-log(data$rent))^2))
        return(list(pred_cv=pred_cv,rmse=rmse))
}

# 计算RMSE
set.seed(123)
rmses=rep(0,50)
for (i in 1:50)
{
        cat(i,"\r")
        pred_cv=pred.cv(dat=data,k=5)
        rmses[i]=pred_cv$rmse
}
mean(rmses)
```

对数线性验证结果 (rmse 值):

```
> mean(rmses)
[1] 0.1522434
```

评估模型结果并进行解读:

从对数线性变化 RMSE 的值可以看到, 对因变量取对数进行预测之后，RMSE 的值变得非常小，这是因为我们对月租金取了对数，使得预测结果和 RMSE 向 0 的方向收缩。不过这个 RMSE 不是实际月租金的 RMSE。转化为实际月租金的 RMSE：

```
reg2=step(lm(log(rent)~.,data=data),direction="both",trace=1,keep=NULL,steps=2000,k=lo

pred.cv<-function(dat,k)
{
        ind=sample(1:k,nrow(dat),replace=T)
        pred_cv=rep(0,nrow(dat))
        sigma <- sum(reg2$residuals^2)/reg2$df.residual
        for (i in 1:k)
        {
                ii=which(ind==i)
                obj=lm(log(rent) ~ bedroom + bathroom + area + subway + region +
                heating,data=dat[-ii,])
                pred_cv[ii]=predict(obj,data[ii,])
        }
        rmse=sqrt(mean((exp(sigma/2)*exp(pred_cv)-data$rent)^2))
        return(list(pred_cv=pred_cv,rmse=rmse))
}

set.seed(123)
rmses=rep(0,50)
for (i in 1:50)
{
        cat(i,"\r")
        pred_cv=pred.cv(dat=data,k=5)
        rmses[i]=pred_cv$rmse
}
mean(rmses)
```

```
> mean(rmses)
[1] 451.8212
```

评估模型结果并进行解读:
转化后的实际月租金 RMSE 上看，其实并没有减少太多，只有略微降幅，说明取对数线性并没有太大程度地降低 RMSE。但是取对数线性模型依旧很好地减少了异方差现象，改善了正态诊断结果，如上一题所示。