短学期作业六

汪利军 *3140105707*

July 10, 2017

Contents

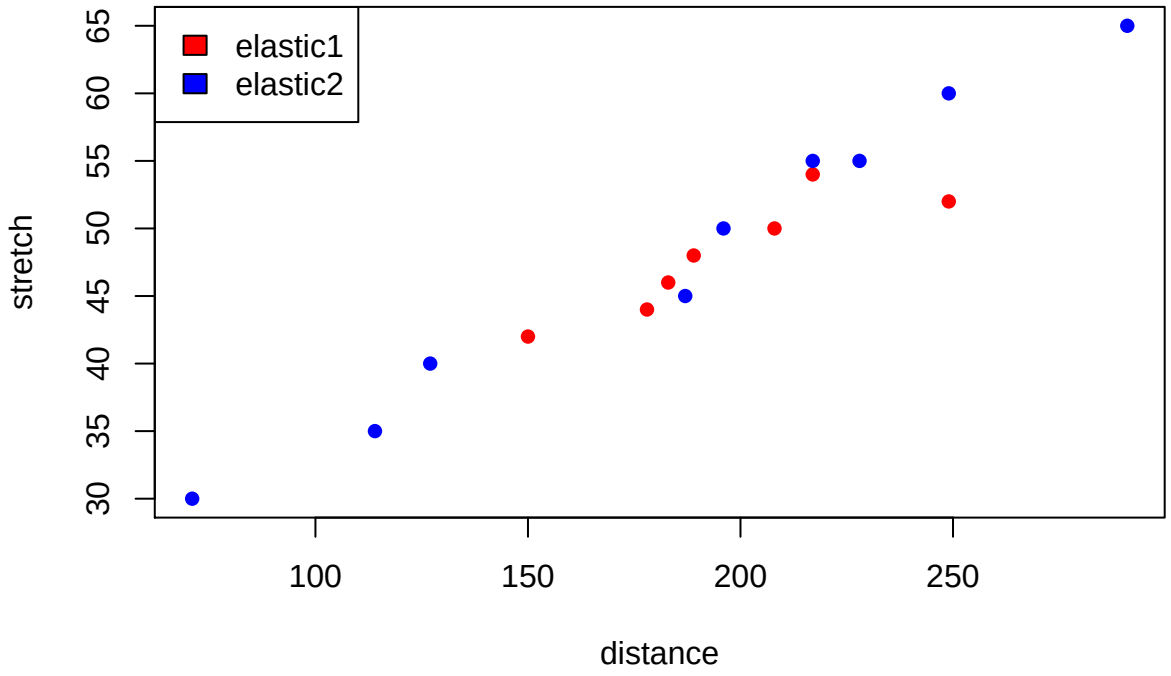
1 MB 5.1

将塑料强度 stretch 看成响应变量，而距离 distance 看成预测变量，则有

```
library(DAAG)
```

```
## Loading required package: lattice
```

```
yrange = c(min(elastic1$stretch, elastic2$stretch),
           max(elastic1$stretch, elastic2$stretch))
xrange = c(min(elastic1$distance, elastic2$distance),
           max(elastic1$distance, elastic2$distance))
with(elastic1, plot(distance, stretch,
                    xlim = xrange, ylim = yrange, col = 'red', pch = 16))
with(elastic2, points(distance, stretch, col = 'blue', pch = 16))
legend("topleft", c("elastic1", "elastic2"), fill = c("red", "blue"))
```



从图象可以看出，虽然两个 elastic 的取值范围不一样，但是整体趋势是一致的。

2 MB 5.2

```
lm.elastic1 <- lm(stretch~distance, data = elastic1)
lm.elastic2 <- lm(stretch~distance, data = elastic2)
```

拟合值为

```
fitted.elastic1 <- predict(lm.elastic1, se.fit = T)
fitted.elastic1$fit
```

```
##        1        2        3        4        5        6        7
## 46.38414 50.51936 47.11388 49.42474 45.77602 42.37054 54.41133
```

```
fitted.elastic2 <- predict(lm.elastic2, se.fit = T)
fitted.elastic2$fit
```

```
##        1        2        3        4        5        6        7        8
## 29.26724 49.87181 38.49809 48.38828 58.60814 53.33338 36.35521 55.14658
##        9
## 65.53128
```

拟合值的标准误差为

```
## elastic1
fitted.elastic1$se.fit
```

```
## [1] 0.8795369 0.9804465 0.8257501 0.8627811 0.9438705 1.4948069 1.6454145
```

```
## elastic2
fitted.elastic2$se.fit
```

```
## [1] 1.1633740 0.5850288 0.7787628 0.5793445 0.7944946 0.6368910 0.8586897
## [8] 0.6823516 1.0787750
```

R^2 为

```
summary(lm.elastic1)$r.squared
```

```
## [1] 0.7992446
```

```
summary(lm.elastic2)$r.squared
```

```
## [1] 0.9807775
```

可见 elastic2 的拟合效果远远好于 elastic1.

使用 rlm() 的结果，

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:DAAG':
##
##     hills
```

```
rlm.elastic1 <- rlm(stretch~distance, data = elastic1)
rlm.elastic2 <- rlm(stretch~distance, data = elastic2)
```

两种拟合的 summary 如下

```
summary(lm.elastic1)
```

```
##
## Call:
## lm(formula = stretch ~ distance, data = elastic1)
##
## Residuals:
##       1       2       3       4       5       6       7
## -0.3841  3.4806  0.8861  0.5753 -1.7760 -0.3705 -2.4113
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.12691    5.41048   4.459  0.00664 **
## distance     0.12162    0.02726   4.462  0.00663 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.121 on 5 degrees of freedom
## Multiple R-squared:  0.7992, Adjusted R-squared:  0.7591
## F-statistic: 19.91 on 1 and 5 DF,  p-value: 0.006631
```

```
summary(lm.elastic2)
```

```
##
## Call:
## lm(formula = stretch ~ distance, data = elastic2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3883 -0.5313  0.1282  1.3919  1.6666
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.563844   1.728140   10.16 1.92e-05 ***
## distance     0.164837   0.008722   18.90 2.89e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.738 on 7 degrees of freedom
## Multiple R-squared:  0.9808, Adjusted R-squared:  0.978
## F-statistic: 357.2 on 1 and 7 DF,  p-value: 2.888e-07
```

```
summary(rlm.elastic1)
```

```
##
```

```
## Call: rlm(formula = stretch ~ distance, data = elastic1)
## Residuals:
##       1       2       3       4       5       6       7
## -0.2207  3.8393  1.0840  0.8822 -1.6413 -0.3967 -1.8689
##
## Coefficients:
##             Value   Std. Error t value
## (Intercept) 25.0146  5.0698     4.9341
## distance     0.1159  0.0255     4.5366
##
## Residual standard error: 1.607 on 5 degrees of freedom
```

```
summary(rlm.elastic2)
```

```
##
## Call: rlm(formula = stretch ~ distance, data = elastic2)
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.49045 -0.63426  0.02595  1.28920  1.56422
##
## Coefficients:
##             Value   Std. Error t value
## (Intercept) 17.6646  1.7257    10.2364
## distance     0.1648  0.0087    18.9266
##
## Residual standard error: 1.911 on 7 degrees of freedom
```

回归系数的比较：两种拟合方式的系数如下，其中括号的是 rlm 的结果，可以看出，lm 和 rlm 这两种方式得到的系数相差不是很大，特别对于 elastic2，结果几乎一致。

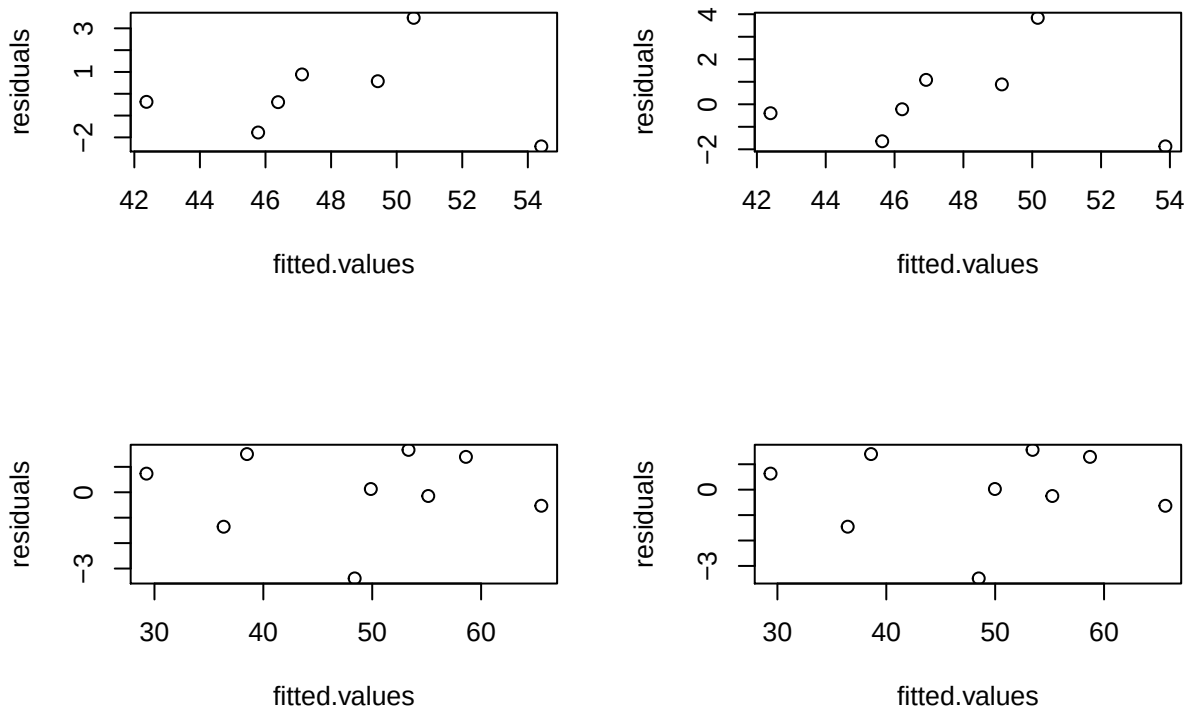
	(Intercept)	distance
elastic1	24.12691(25.0146)	0.12162(0.1159)
elastic2	17.563844(17.6646)	0.164837(0.1648)

系数的标准误差的比较：两种拟合方式的系数的标准误差如下，可见这两种方式的系数标准误差相差不大，但使用 rlm 的系数的标准误差都略小于 lm 的系数的标准误差。

	(Intercept) SE	distance SE
elastic1	5.41048(5.0698)	0.02726(0.0255)
elastic2	1.728140(1.7257)	0.008722(0.0087)

残差图的比较

```
par(mfrow = c(2,2))
plot(lm.elastic1$fitted.values, lm.elastic1$residuals,
     xlab = "fitted.values", ylab = "residuals")
plot(rlm.elastic1$fitted.values, rlm.elastic1$residuals,
     xlab = "fitted.values", ylab = "residuals")
plot(lm.elastic2$fitted.values, lm.elastic2$residuals,
     xlab = "fitted.values", ylab = "residuals")
plot(rlm.elastic2$fitted.values, rlm.elastic2$residuals,
     xlab = "fitted.values", ylab = "residuals")
```



第一排是 elastic1 的残差图，第二排为 elastic2 的残差图，可以观察发现对于每个数据集，两种回归的残差图模式基本一致。

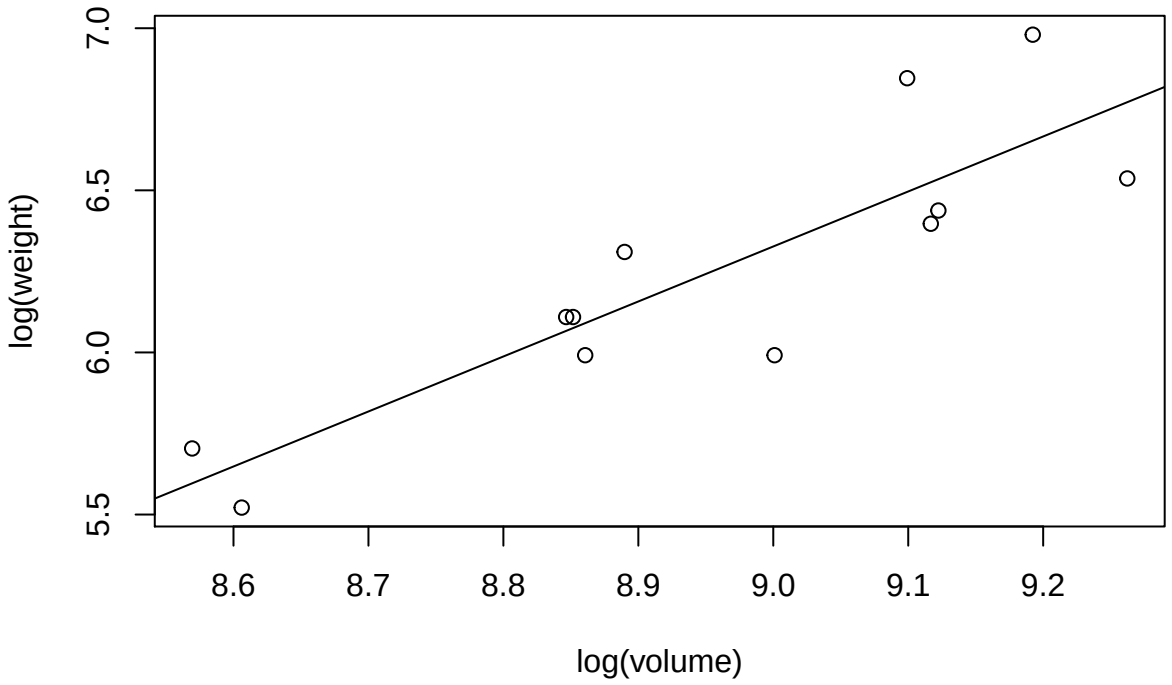
3 MB 5.4

3.1 (a)

首先在 oddbooks 中增加一列 volume

```
oddbooks <- within(oddbooks, volume <- thick*height*breadth)
```

```
with(oddbooks, plot(log(volume), log(weight)))
lm.oddbooks <- lm(log(weight)~log(volume), data = oddbooks)
abline(lm.oddbooks)
```

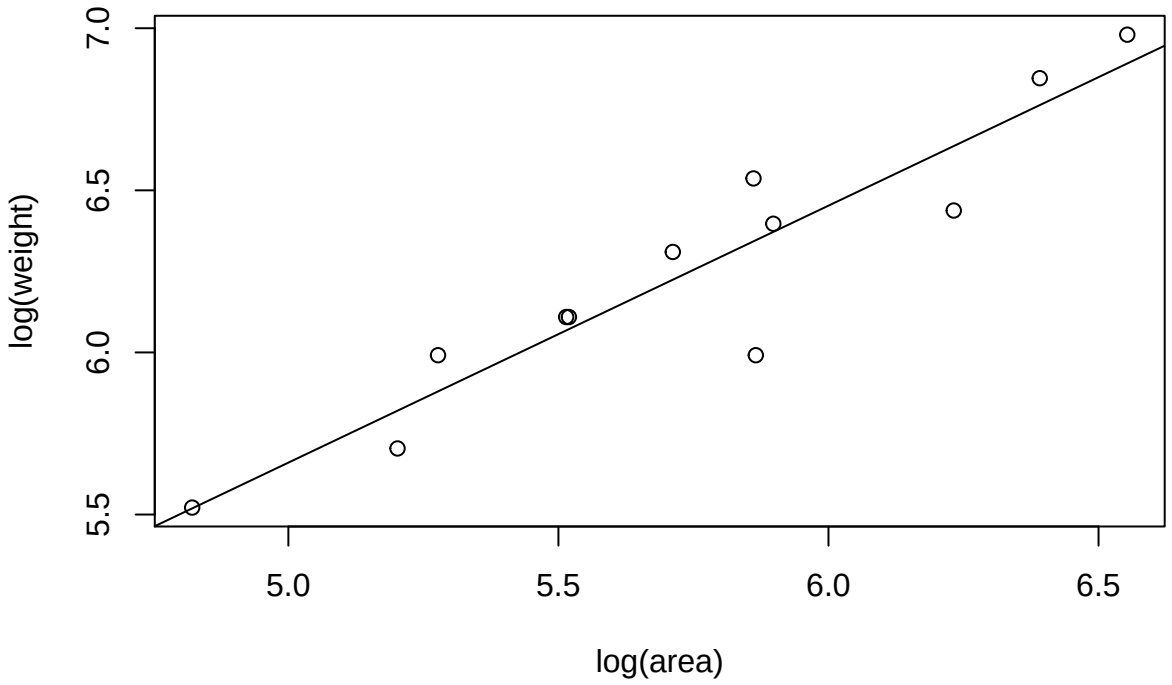


3.2 (b)

首先在 oddbooks 中增加一列 area

```
oddbooks <- within(oddbooks, area <- height*breadth)
```

```
with(oddbooks, plot(log(area), log(weight)))
lm.oddbooks2 <- lm(log(weight)~log(area), data = oddbooks)
abline(lm.oddbooks2)
```



3.3 (c)

(a) 的回归为

```
summary(lm.oddbooks)
```

```
##
## Call:
## lm(formula = log(weight) ~ log(volume), data = oddbooks)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33694 -0.13018 -0.03118  0.12317  0.35076
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -8.942      2.731  -3.274  0.00837 **
## log(volume)    1.696      0.305   5.562  0.00024 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2228 on 10 degrees of freedom
## Multiple R-squared:  0.7557, Adjusted R-squared:  0.7313
## F-statistic: 30.94 on 1 and 10 DF,  p-value: 0.00024
```

(b) 的回归为

```r
summary(lm.oddbooks2)
```

```
##
## Call:
## lm(formula = log(weight) ~ log(area), data = oddbooks)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35463 -0.02720  0.03948  0.08644  0.19402
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.69675    0.54648   3.105   0.0112 *
## log(area)    0.79267    0.09491   8.352 8.06e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1597 on 10 degrees of freedom
## Multiple R-squared:  0.8746, Adjusted R-squared:  0.8621
## F-statistic: 69.75 on 1 and 10 DF,  p-value: 8.063e-06
```

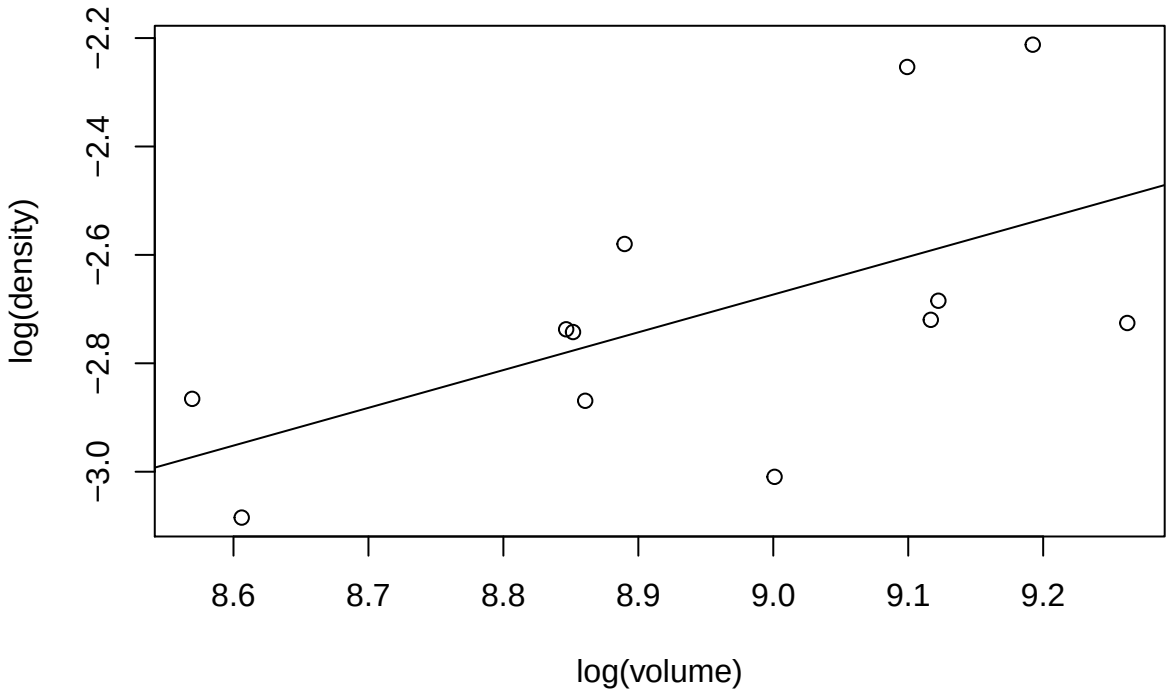
如果从 R^2 来看的话，(b) 的拟合更好，因 $0.8746 > 0.7557$。

3.4 (d)

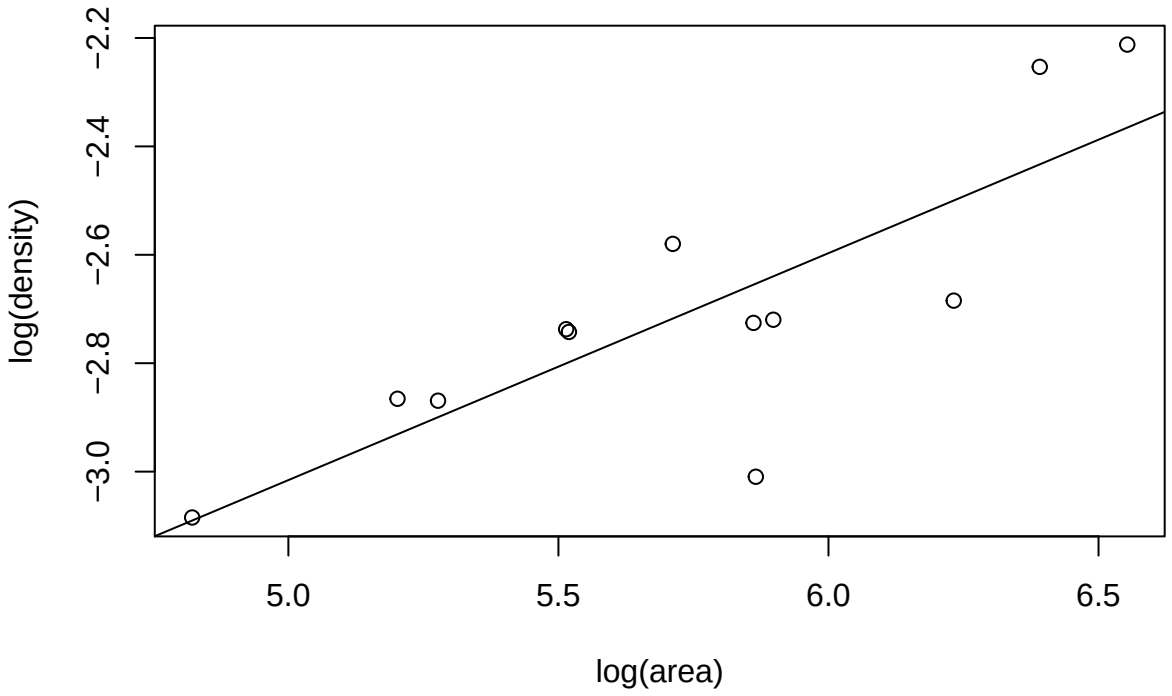
首先在 oddbooks 中增加一列 density

```r
oddbooks <- within(oddbooks, density <- weight/volume)
```

```r
with(oddbooks, plot(log(volume), log(density)))
lm.oddbooks.den.a <- lm(log(density)~log(volume), data = oddbooks)
abline(lm.oddbooks.den.a)
```




```
with(oddbooks, plot(log(area), log(density)))
lm.oddbooks.den.b <- lm(log(density)~log(area), data = oddbooks)
abline(lm.oddbooks.den.b)
```



回归结果为

```
summary(lm.oddbooks.den.a)
```

```
##
```

```
## Call:
## lm(formula = log(density) ~ log(volume), data = oddbooks)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33694 -0.13018 -0.03118  0.12317  0.35076
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -8.9420     2.7311  -3.274  0.00837 **
## log(volume)   0.6965     0.3050   2.284  0.04551 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2228 on 10 degrees of freedom
## Multiple R-squared:  0.3427, Adjusted R-squared:  0.277
## F-statistic: 5.215 on 1 and 10 DF,  p-value: 0.04551
```

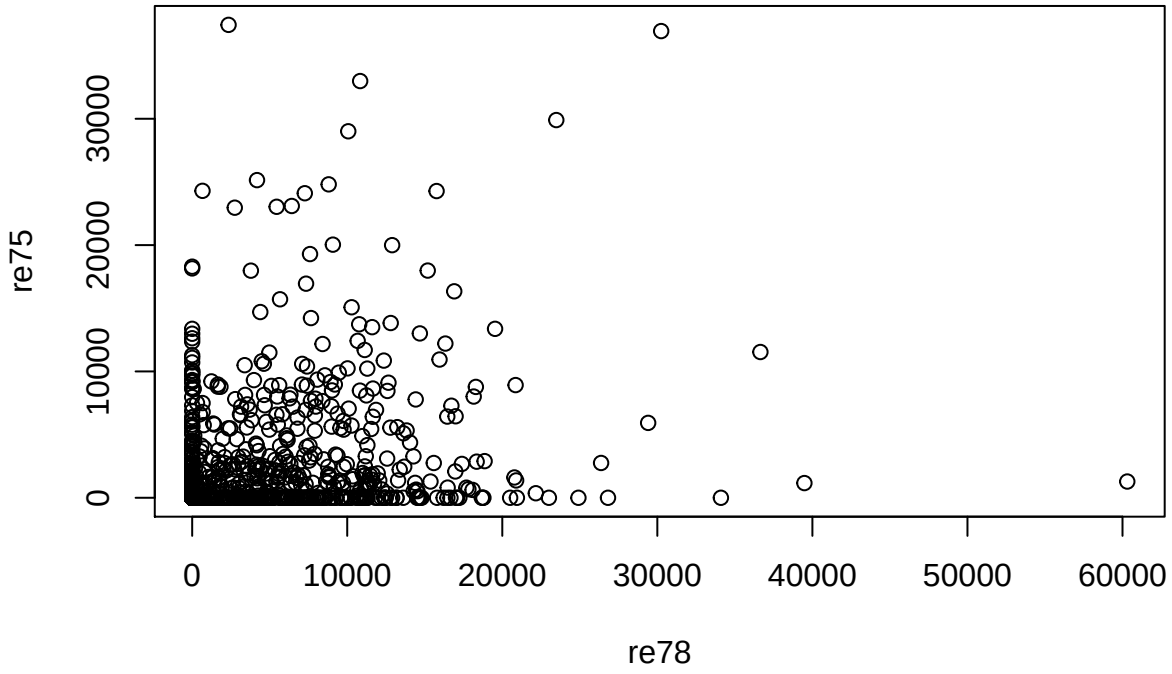
```
summary(lm.oddbooks.den.b)
```

```
##
## Call:
## lm(formula = log(density) ~ log(area), data = oddbooks)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35615 -0.07288  0.04320  0.08355  0.17988
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.10912    0.55138  -9.266 3.18e-06 ***
## log(area)    0.41869    0.09576   4.372  0.00139 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1611 on 10 degrees of freedom
## Multiple R-squared:  0.6565, Adjusted R-squared:  0.6222
## F-statistic: 19.12 on 1 and 10 DF,  p-value: 0.001394
```

从 R^2 来看，两者的拟合效果都不好，R^2 都不足 0.7；尽管如此，(b) 的拟合效果更好。

4 MB 5.9

```
with(nswdemo, plot(re78, re75))
```



从图象可以看出，

1. 数据点的波动极差较大，大部分点较小，然而也存在值很大的点，造成数据点堆聚在原点附近的小区域中。
2. 存在很多值为 0 的点，导致大部分点落在左边轴上。

4.1 (a)

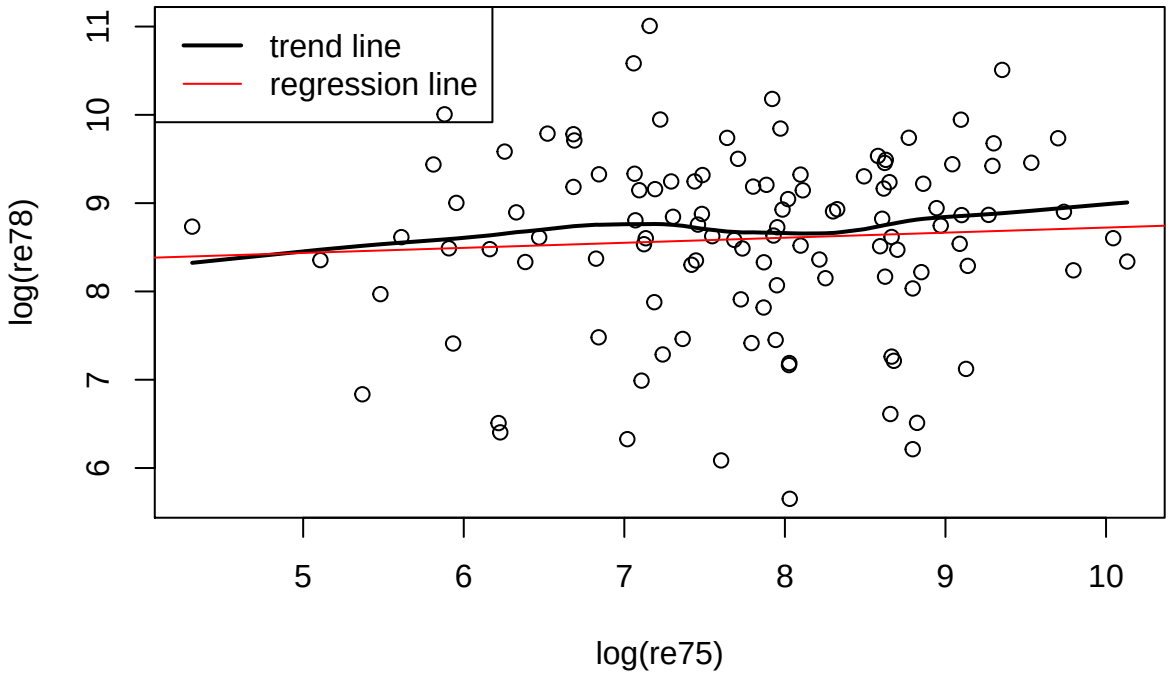
去除值为 0 的数据

```
nsw74demo <- within(nsw74demo, {
  re75[abs(re75) < 1e-10] <-NA
  re78[abs(re78) < 1e-10] <-NA
})
nsw74demo.omit.na <- nsw74demo[complete.cases(nsw74demo),]
```

作图

```
with(nsw74demo.omit.na, plot(log(re75), log(re78)))
## fitting a smooth trend curve
with(nsw74demo.omit.na, lines(lowess(log(re75), log(re78)), lwd = 2))
## regression line
lm.nsw74demo <- lm(log(re78)~log(re75), data = nsw74demo.omit.na)
```

```r
abline(lm.nsw74demo, col = "red")
legend("topleft", c("trend line", "regression line"),
       lty = c(1, 1), lwd = c(2, 1), col = c("black", "red"))
```

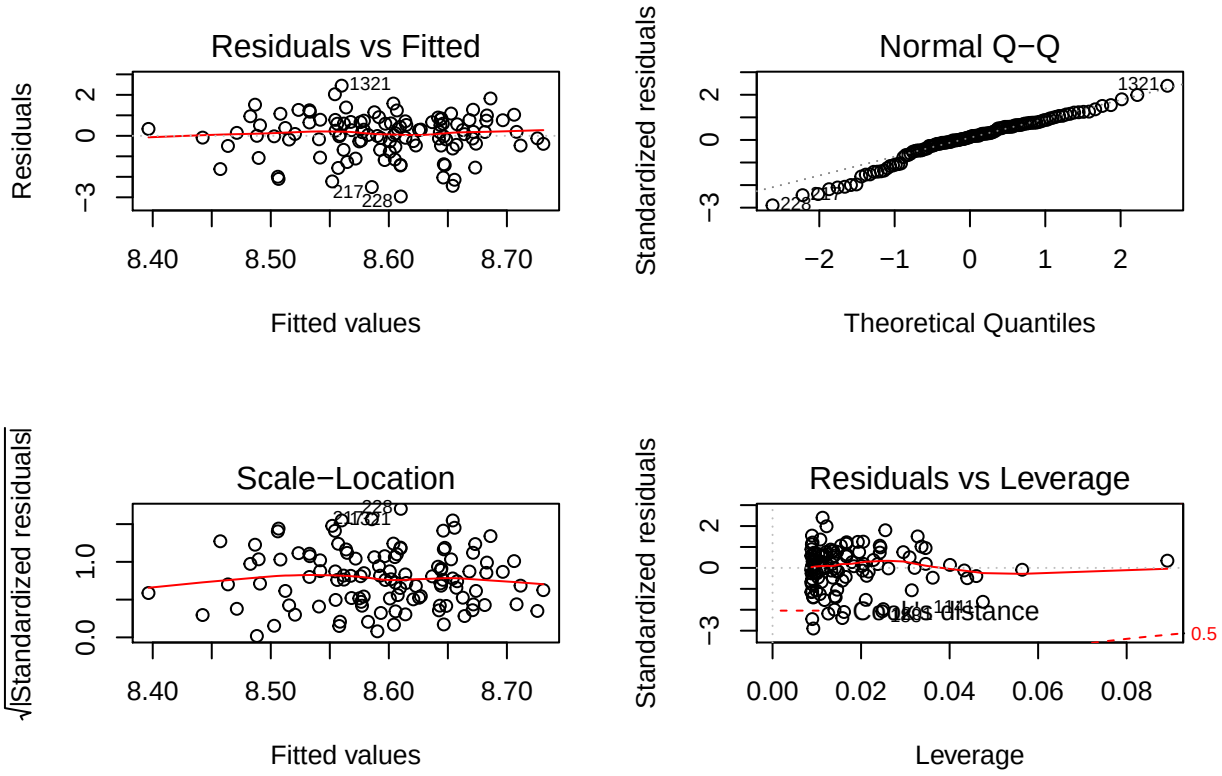


从图中可以看出，回归曲线和趋势曲线大致一致，所以回归曲线大致上能够描述两者之间的关系，但是斜率偏小。

4.2 (b)

四张模型诊断曲线如下

```r
par(mfrow = c(2,2))
plot(lm.nsw74demo)
```



从 qq 图可以看出，除了较小的值偏离理论的分位数，其他值都落在 y=x 这条直线上，表明近似符合正态性假设，考虑到较小的值偏小可以进一步考虑采用新的数据变换来弥补这一不足；从其他残差图中看没有特别明显的形状，也就是符合不相关性的假设；而从残差杠杆图可以看出，有个别的观测有较大的杠杆值，再进一步的研究中可以考虑删去这些离群点重新拟合。

5 MB 5.10

```
simY <- function(x)
{
  nlen = length(x)
  return(2+3*x+rnorm(nlen))
}
```

当对 x 均匀采样时

```
set.seed(123)
x1 = runif(10, -1, 1)
y1 = simY(x1)
```

当 $x_i \in \{-1, 1\}, i = 1, \ldots, 10$ 时

```r
set.seed(123)
x2 = rep(c(-1, 1), 5)
y2 = simY(x2)
```

进行线性回归有

```r
lm.1 <- lm(y1 ~ x1)
lm.2 <- lm(y2 ~ x2)
```

两个回归的结果为

```r
summary(lm.1)
```

```
##
## Call:
## lm(formula = y1 ~ x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.62155 -0.33471  0.05238  0.55227  1.19742
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.236      0.284   7.873  4.9e-05 ***
## x1             2.336      0.489   4.778   0.0014 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8648 on 8 degrees of freedom
## Multiple R-squared:  0.7405, Adjusted R-squared:  0.708
## F-statistic: 22.82 on 1 and 8 DF,  p-value: 0.001395
```

```r
summary(lm.2)
```

```
##
## Call:
## lm(formula = y2 ~ x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2340 -0.6592 -0.1251  0.2358  1.7461
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.0746     0.3177    6.53 0.000182 ***
## x2            2.8943     0.3177    9.11  1.7e-05 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.005 on 8 degrees of freedom
## Multiple R-squared:  0.9121, Adjusted R-squared:  0.9011
## F-statistic: 82.99 on 1 and 8 DF,  p-value: 1.695e-05
```

由上述结果我们有

	slope	std. err	noise std. deviation
lm.1	2.336	0.8648	22.82
lm.2	2.8943	0.3177	1.005

可见对于第二种拟合效果更好一点。这两种设计的优缺点分别是

1. 设计一依赖于每次随机的数据，所以改变 seed 会有不同的拟合结果，由于数据量较小，所以有较大可能使得拟合的模型很好；但这种随机性恰恰是设计二所不具备的
2. 设计二的 x 是固定的，得到的回归模型更加稳定，不会随着随机种子的改变而改变。

6 MB 5.11

6.1 (a)

```
library(DAAG)
e1.lm <- lm(distance~ stretch, data = elastic1)
elastic1$newdistance <-
  cbind(rep(1, 7), elastic1$stretch)%*%coef(e1.lm) +
  rnorm(7, sd=summary(e1.lm)$sigma)
```

6.2 (b)

simulate 与 (a) 中的代码有相同的效果

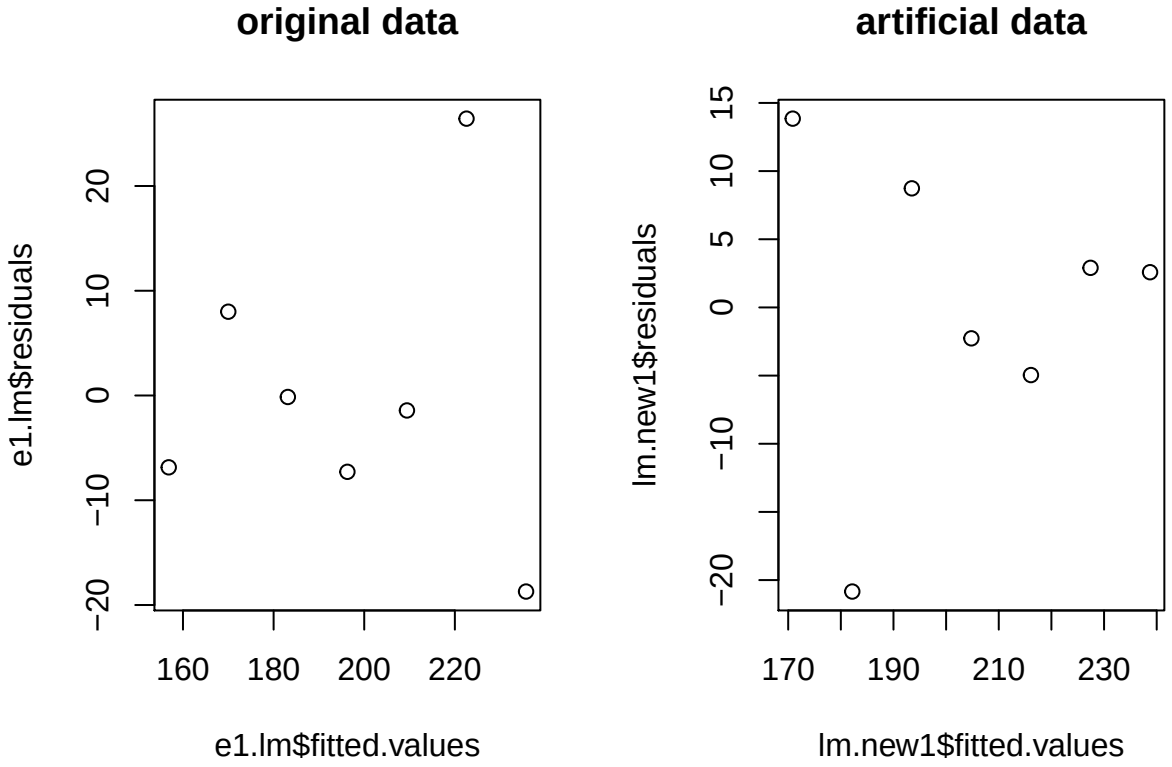
```
simulate(e1.lm)
```

```
##      sim_1
## 1 152.4881
## 2 246.6467
## 3 188.9161
## 4 192.7838
## 5 166.6023
## 6 140.8643
```

```
## 7 211.2098
```

6.3 (c)

```
par(mfrow = c(1,2))
lm.new1 <- lm(newdistance ~ stretch, data = elastic1)
plot(e1.lm$fitted.values, e1.lm$residuals, main = "original data")
plot(lm.new1$fitted.values, lm.new1$residuals, main = "artificial data")
```



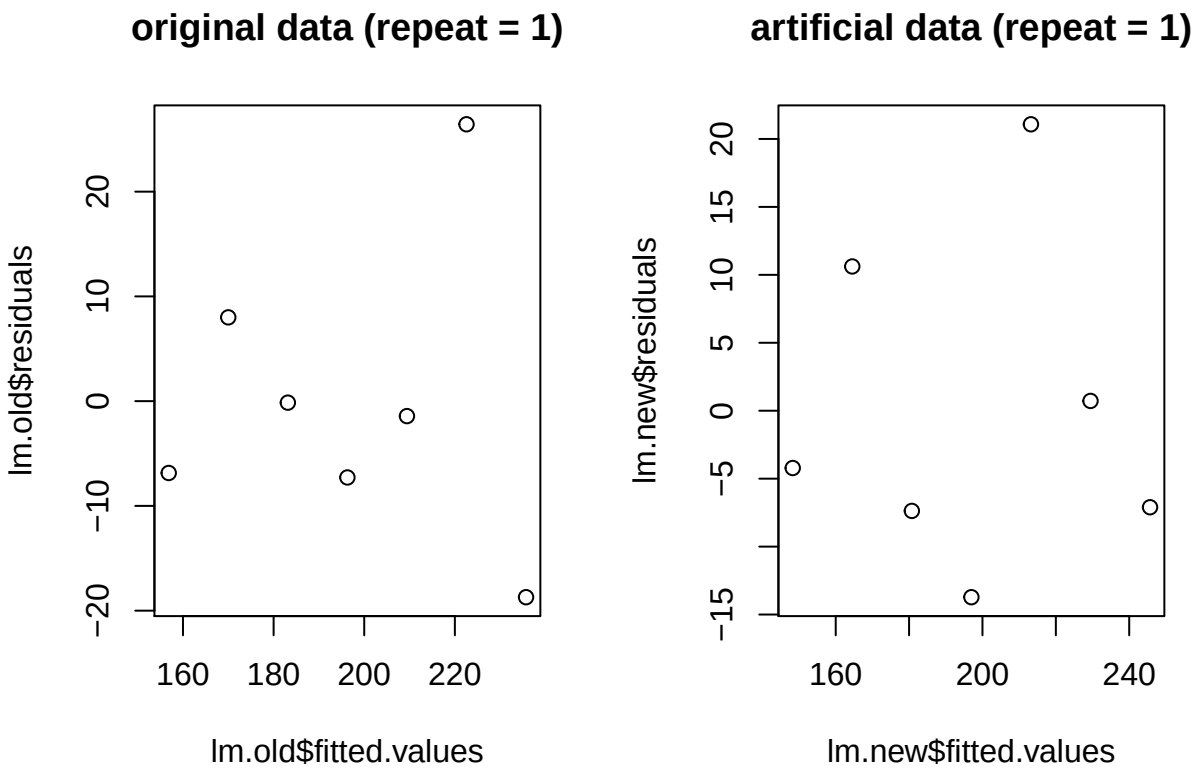
为了重复上述操作，编写下面的 repeatPlots 函数

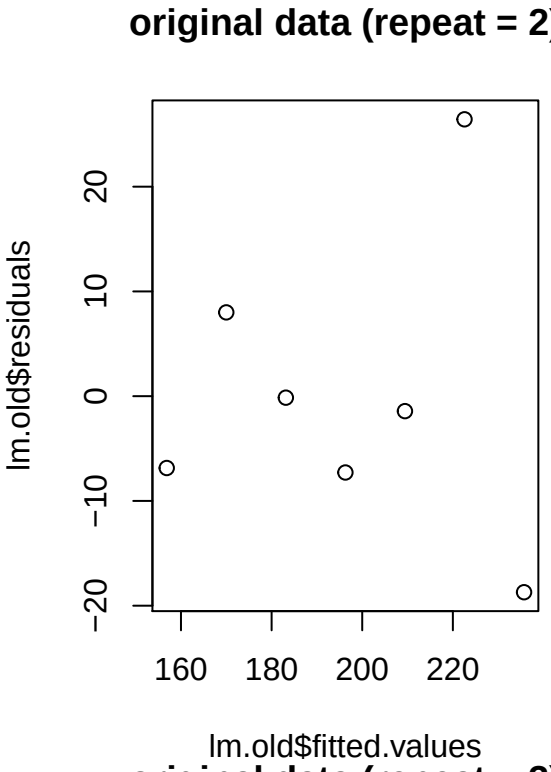
```
repeatPlots <- function(n, dataset)
{
  lm.old <- lm(distance~ stretch, data = dataset)
  for (i in 1:n)
  {
    dataset$newdistance <-
    cbind(rep(1, nrow(dataset)), dataset$stretch)%*%coef(lm.old) +
      rnorm(nrow(dataset), sd=summary(lm.old)$sigma)

    lm.new <- lm(newdistance~stretch, dataset)
    par(mfrow = c(1,2))
    plot(lm.old$fitted.values, lm.old$residuals,
         main = paste0("original data (repeat = ", i, ")"))
```

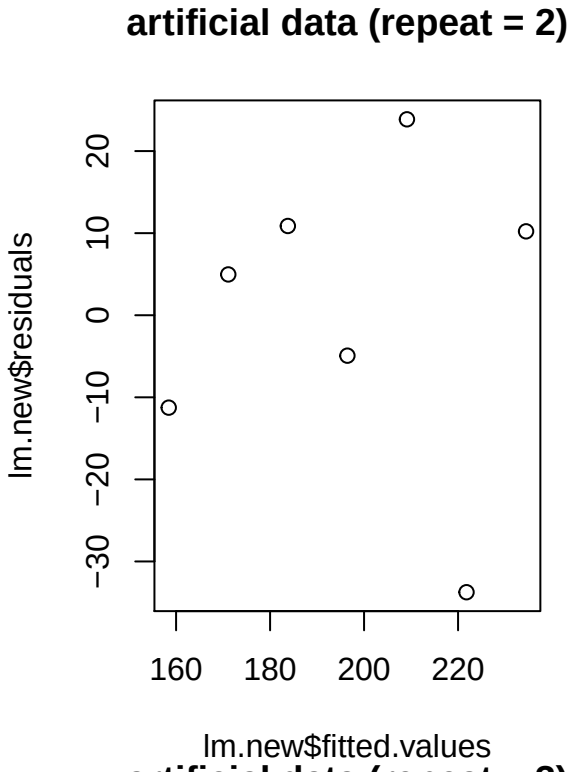
```
    plot(lm.new$fitted.values, lm.new$residuals,
         main = paste0("artificial data (repeat = ", i, ")"))
  }
}
```

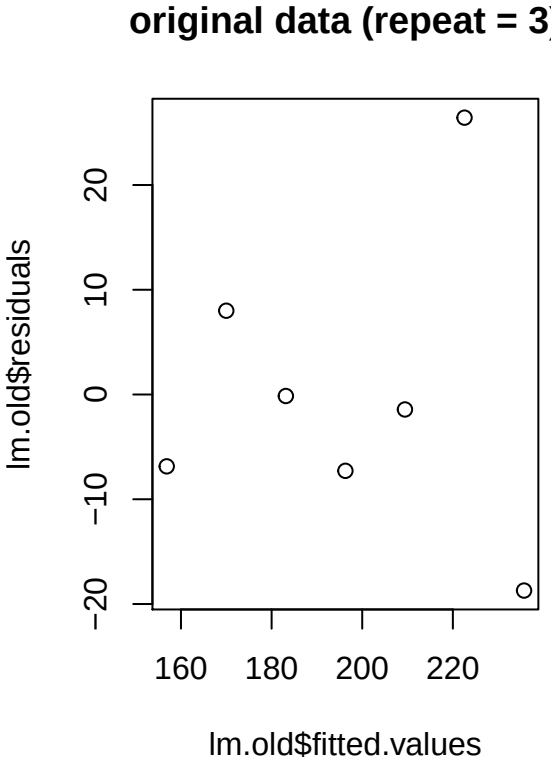
重复 5 次，

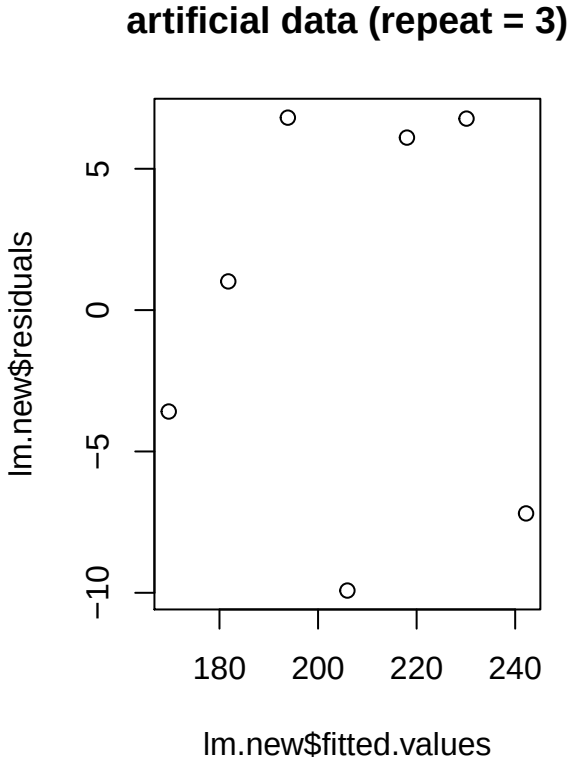
```
set.seed(1)
repeatPlots(5, elastic1)
```

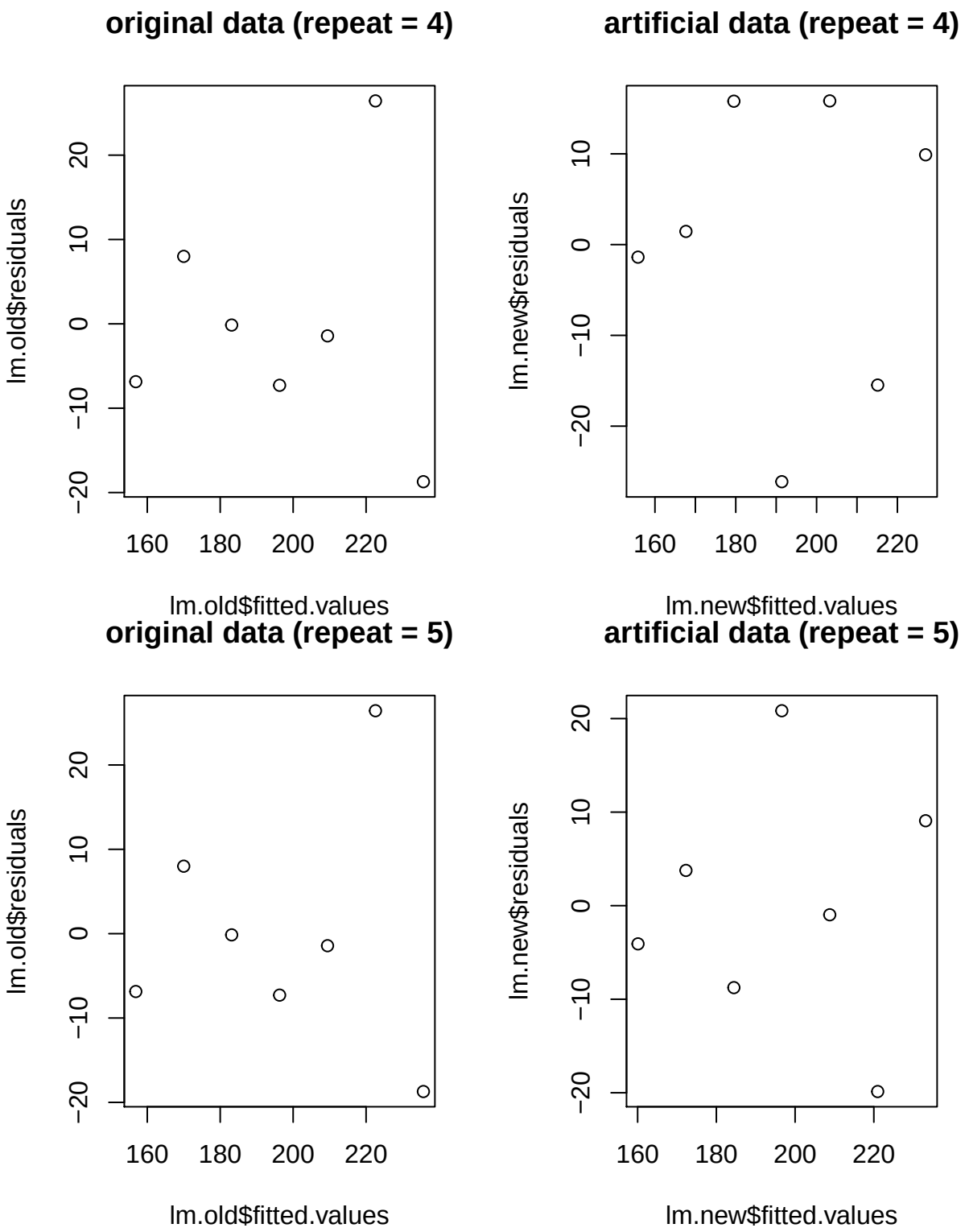



original data (repeat = 2)
lm.old$residuals
lm.old$fitted.values
−20
−10
0
10
20
160
180
200
220


artificial data (repeat = 2)
lm.new$residuals
lm.new$fitted.values
−30
−20
−10
0
10
20
160
180
200
220


original data (repeat = 3)
lm.old$residuals
lm.old$fitted.values
−20
−10
0
10
20
160
180
200
220

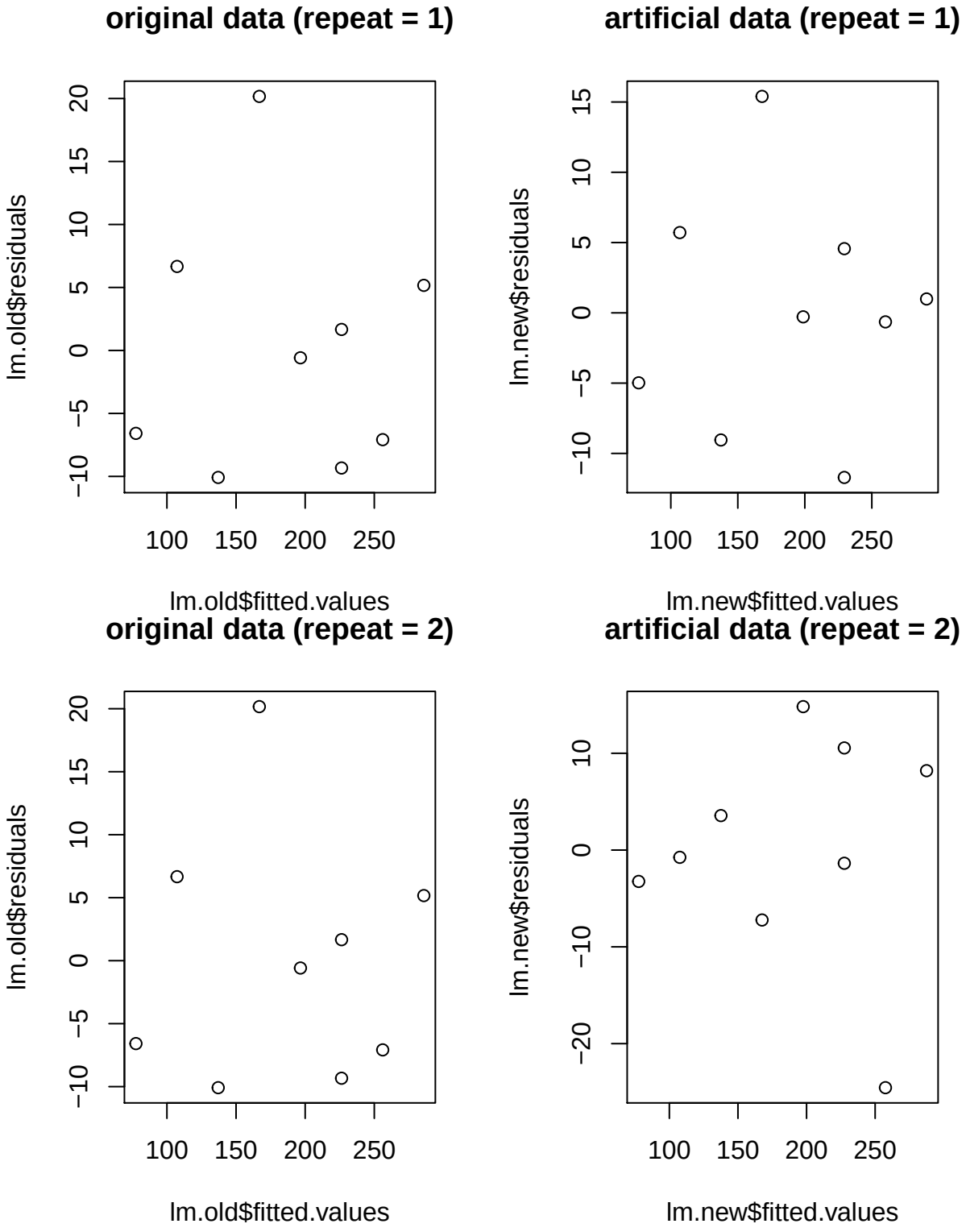

artificial data (repeat = 3)
lm.new$residuals
lm.new$fitted.values
−10
−5
0
5
180
200
220
240

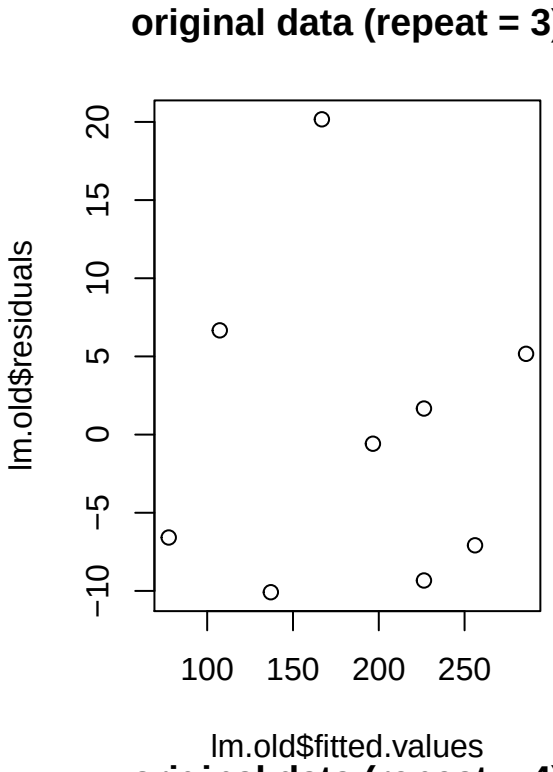


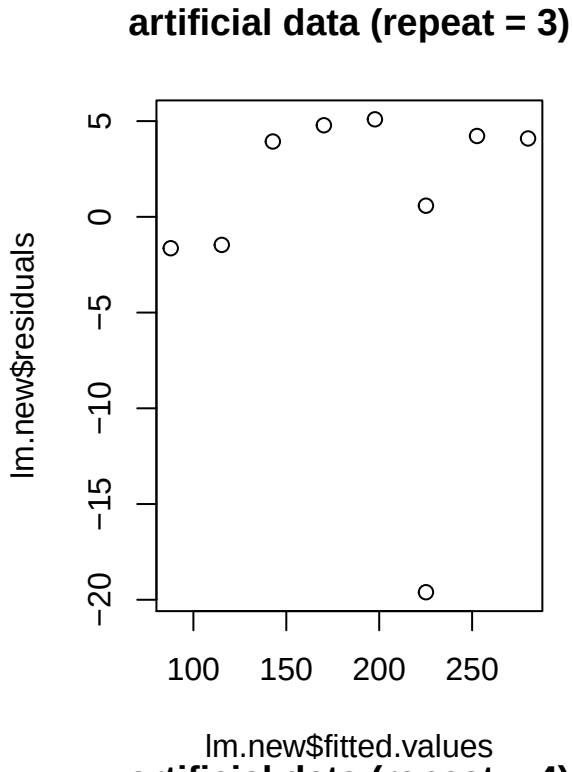
原始残差图中的异常点不一定是人造数据中的异常点，比如在 repeat=1 中，第 6 个点是异常点，然而在人造数据中第六个的的残差已经很小了，所以原始残差图和人造残差图的异常点不一定一致。

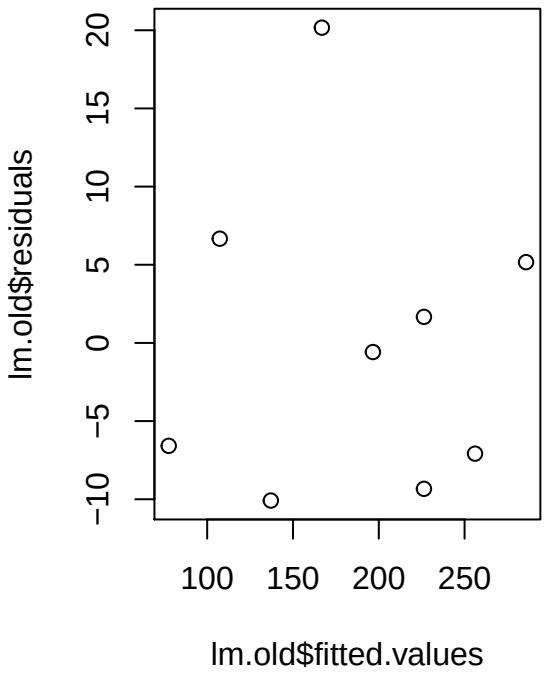
对于 elastic2 我们也有重复 5 次后的结果

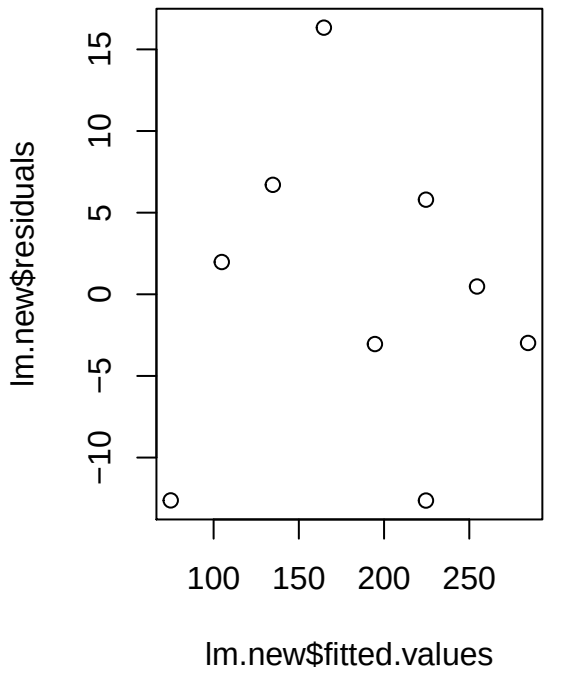
```
set.seed(1)
repeatPlots(5, elastic2)
```

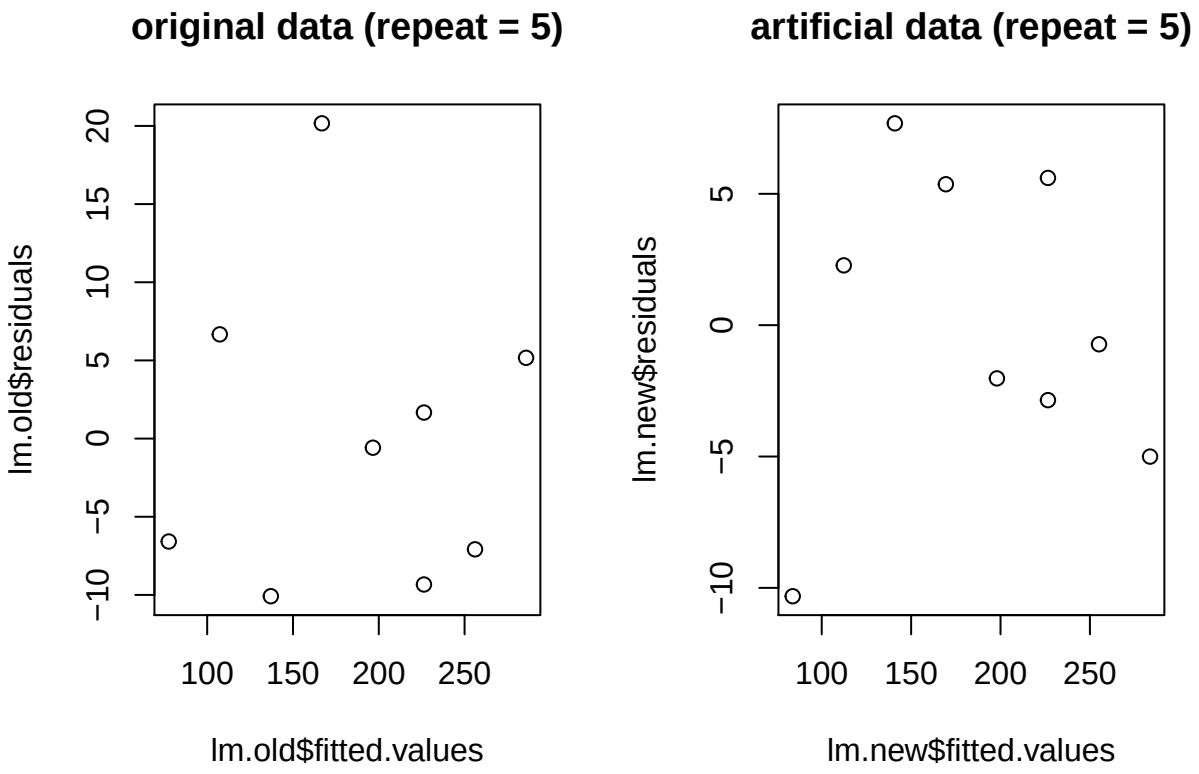



original data (repeat = 3)
lm.old$residuals
lm.old$fitted.values
−10
−5
0
5
10
15
20
100
150
200
250


artificial data (repeat = 3)
lm.new$residuals
lm.new$fitted.values
−20
−15
−10
−5
0
5
100
150
200
250


original data (repeat = 4)
lm.old$residuals
lm.old$fitted.values
−10
−5
0
5
10
15
20
100
150
200
250


artificial data (repeat = 4)
lm.new$residuals
lm.new$fitted.values
−10
−5
0
5
10
15
100
150
200
250



7 MDL Chapter 14 Worksheet A: Study of intima media

7.1 Problem 14.1

```
library(XLConnect)
```

```
## Loading required package: XLConnectJars
```

```
## XLConnect 0.2-13 by Mirai Solutions GmbH [aut],
##   Martin Studer [cre],
##   The Apache Software Foundation [ctb, cph] (Apache POI),
##   Graph Builder [ctb, cph] (Curvesapi Java library)
```

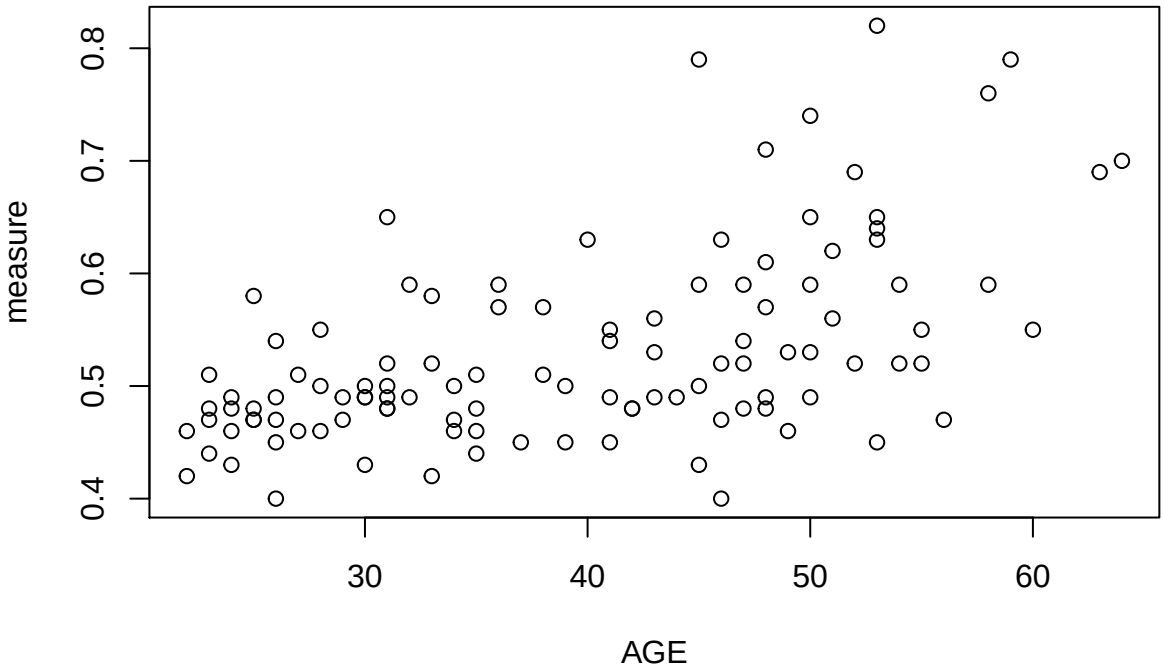
```
## http://www.mirai-solutions.com ,
## http://miraisolutions.wordpress.com
```

```
tmp = tempfile(fileext = ".xls")
download.file(url = "http://biostatisticien.eu/springeR/Intima_Media_Thickness.xls",
              destfile = tmp, mode = "wb")
connect =  loadWorkbook(tmp)
data = readWorksheet(connect, 1)
head(data)
```

```
##   GENDER AGE height weight tobacco packyear SPORT measure alcohol
## 1      1  33    170     70       1        1     0    0.52       1
## 2      2  33    177     67       2       20     0    0.42       1
## 3      2  53    164     63       1       30     0    0.65       0
## 4      2  42    169     76       1       26     1    0.48       1
## 5      2  53    152     54       0       NA     0    0.45       1
## 6      2  50    162     53       2       10     0    0.49       1
```

7.2 Problem 14.2

```
with(data, plot(measure ~ AGE))
```



从散点图可以看出，随着 AGE 的增大，measure 也增大。

7.3 Problem 14.3

采用相关系数来衡量变量直接的联系，相关系数越大，联系越强。得到如下的相关系数矩阵，进而可以判断两者之间的相关性。

```
cor(data)
```

```
##                GENDER         AGE      height      weight     tobacco
## GENDER     1.00000000  0.24584709 -0.64366517 -0.51540287 -0.11746665
## AGE        0.24584709  1.00000000 -0.34641559 -0.04446022  0.27561732
## height    -0.64366517 -0.34641559  1.00000000  0.56887235  0.06665331
## weight    -0.51540287 -0.04446022  0.56887235  1.00000000  0.15865637
```

```
## tobacco  -0.11746665  0.27561732  0.06665331  0.15865637  1.00000000
## packyear          NA          NA          NA          NA          NA
## SPORT    -0.08752204 -0.17335604  0.13266979 -0.02764585  0.05055093
## measure   0.01992217  0.55236686 -0.06865902  0.22952003  0.20922141
## alcohol  -0.25730613  0.24494641  0.16644351  0.22473789  0.19008078
##          packyear       SPORT     measure     alcohol
## GENDER         NA -0.08752204  0.01992217 -0.25730613
## AGE            NA -0.17335604  0.55236686  0.24494641
## height         NA  0.13266979 -0.06865902  0.16644351
## weight         NA -0.02764585  0.22952003  0.22473789
## tobacco        NA  0.05055093  0.20922141  0.19008078
## packyear        1          NA          NA          NA
## SPORT          NA  1.00000000 -0.09825353  0.03454682
## measure        NA -0.09825353  1.00000000  0.23531880
## alcohol        NA  0.03454682  0.23531880  1.00000000
```

7.4 Problem 14.4

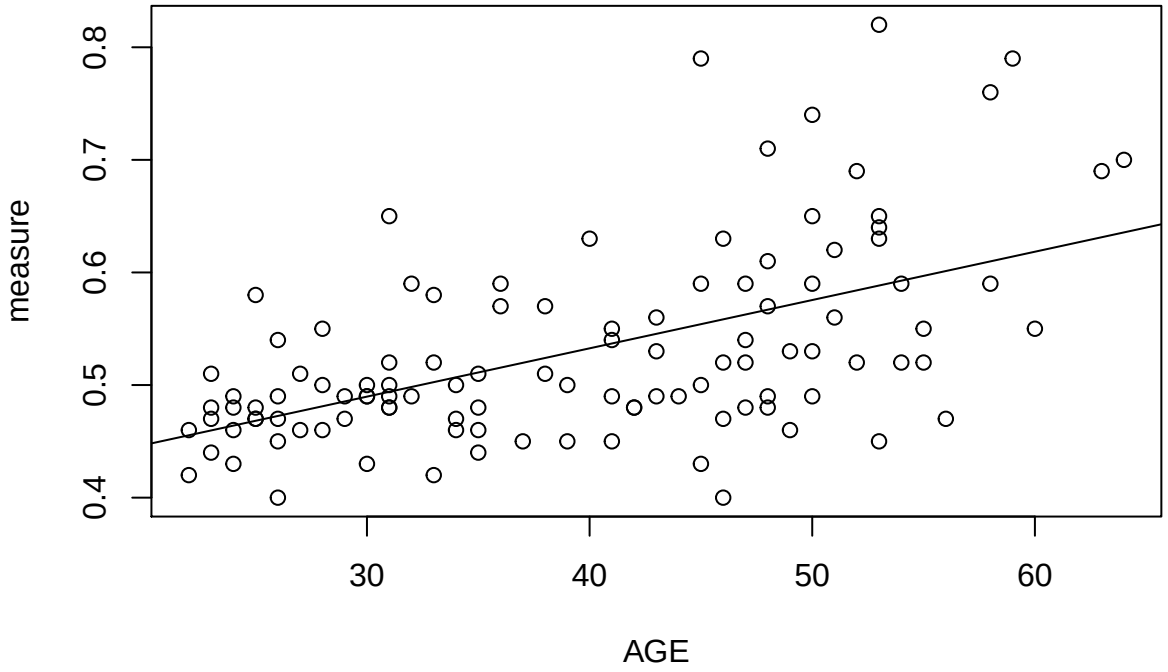
```
lm.data <- lm(measure~AGE, data = data)
summary(lm.data)
```

```
##
## Call:
## lm(formula = measure ~ AGE, data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.158351 -0.046944 -0.003323  0.035243  0.235939
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.3610255  0.0255330  14.140  < 2e-16 ***
## AGE         0.0042897  0.0006229   6.886 3.95e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07257 on 108 degrees of freedom
## Multiple R-squared:  0.3051, Adjusted R-squared:  0.2987
## F-statistic: 47.42 on 1 and 108 DF,  p-value: 3.953e-10
```

参数估计如上所返回的结果所示。

将回归曲线添加到散点图中

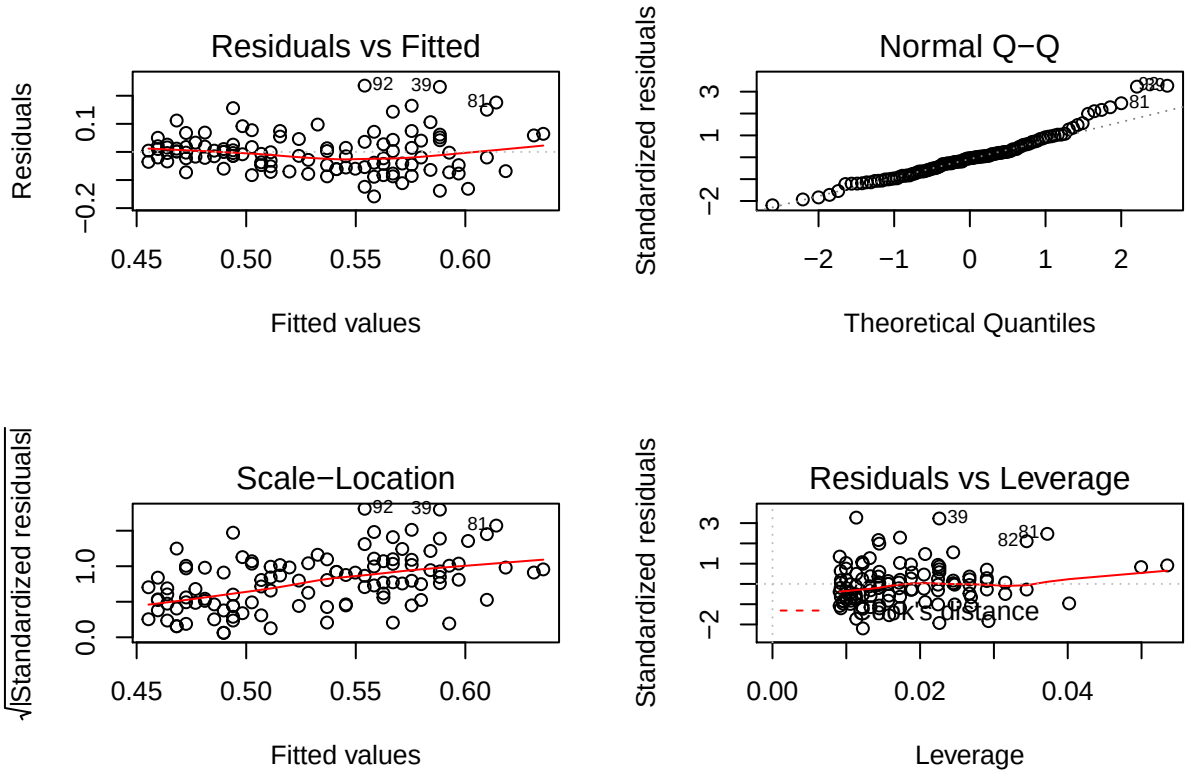
```r
with(data, plot(measure ~ AGE))
abline(lm.data)
```



7.5 Problem 14.5

回归模型的诊断图如下，

```r
par(mfrow=c(2,2))
plot(lm.data)
```



从 QQ 图可以看出残差有重尾现象，理论分位数高的地方，标准残差的分位数也较高，但在一定精度范围内，也可以认定满足残差正态性的假设；另外从残差图中也可以观察出它满足线性的假设，因为其与拟合值不存在明显的线性关系；从 Scale-Location 也可以看出它不太满足方差齐性的假设，因为它没有随机散布在水平直线附近，而是一条有明显斜率的曲线附近，这表明需要进一步优化模型。

7.6 Problem 14.6

```
predict(lm.data, data.frame(AGE=33), interval = "prediction")
```

```
##         fit       lwr       upr
## 1 0.5025848 0.3578677 0.647302
```

从返回结果可以看出，预测值为 0.5025848，prediction interval 为 $[0.3578677, 0.647302]$

7.7 Problem 14.7

```
predict(lm.data, data.frame(AGE=33), interval = "confidence")
```

```
##         fit       lwr       upr
## 1 0.5025848 0.4867222 0.5184474
```

从返回结果可以看出，预测值为 0.5025848，confidence interval 为 $[0.4867222, 0.5184474]$

7.8 Problem 14.8

新模型如下

```
lm.data2 <- lm(measure ~ AGE + I(AGE^2) , data = data)
summary(lm.data2)
```

```
##
## Call:
## lm(formula = measure ~ AGE + I(AGE^2), data = data)
##
## Residuals:
##       Min         1Q     Median         3Q        Max
## -0.148064 -0.042509 -0.009462  0.029712  0.247419
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.305e-01  9.295e-02   5.707 1.04e-07 ***
## AGE         -4.832e-03  4.855e-03  -0.995   0.3219
## I(AGE^2)     1.134e-04  5.984e-05   1.894   0.0609 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07172 on 107 degrees of freedom
## Multiple R-squared:  0.3277, Adjusted R-squared:  0.3151
## F-statistic: 26.07 on 2 and 107 DF,  p-value: 5.974e-10
```

此时 R^2 为 0.3277，略高于原来的 0.3051，因此在 R^2 意义下，该模型更好。