短学期作业七

汪利军 3140105707

July 10, 2017

Contents

1 MB 6.1

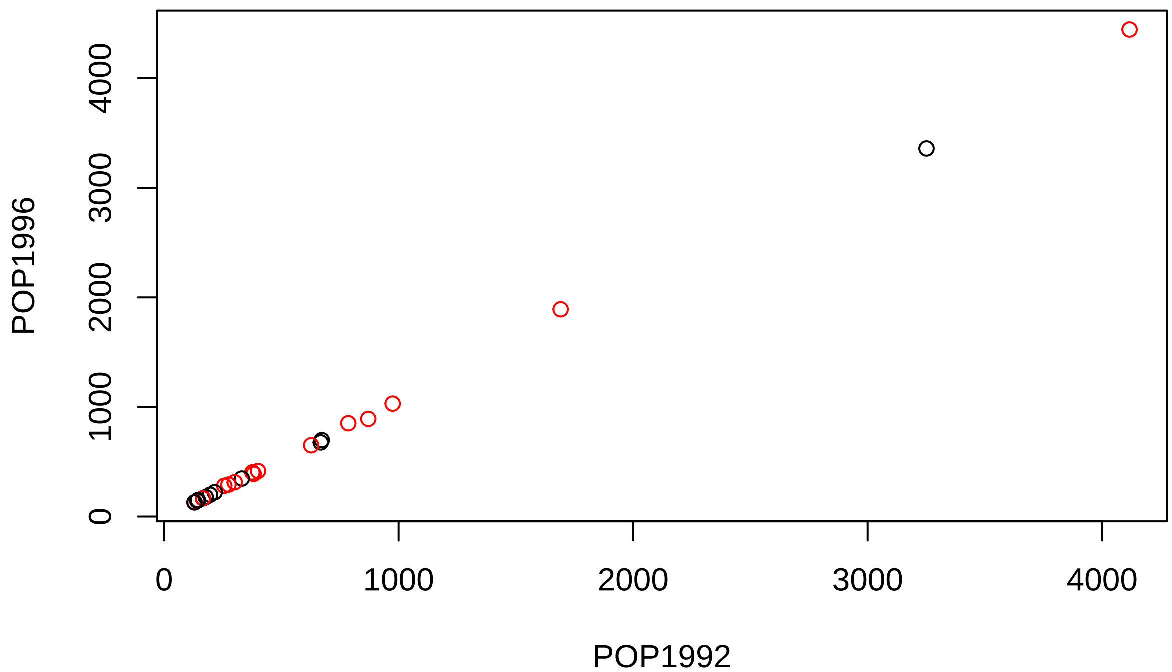
```
library(DAAG)
```

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

```
cities$have <- factor((cities$REGION == "ON") |
                       (cities$REGION == "WEST"))
```

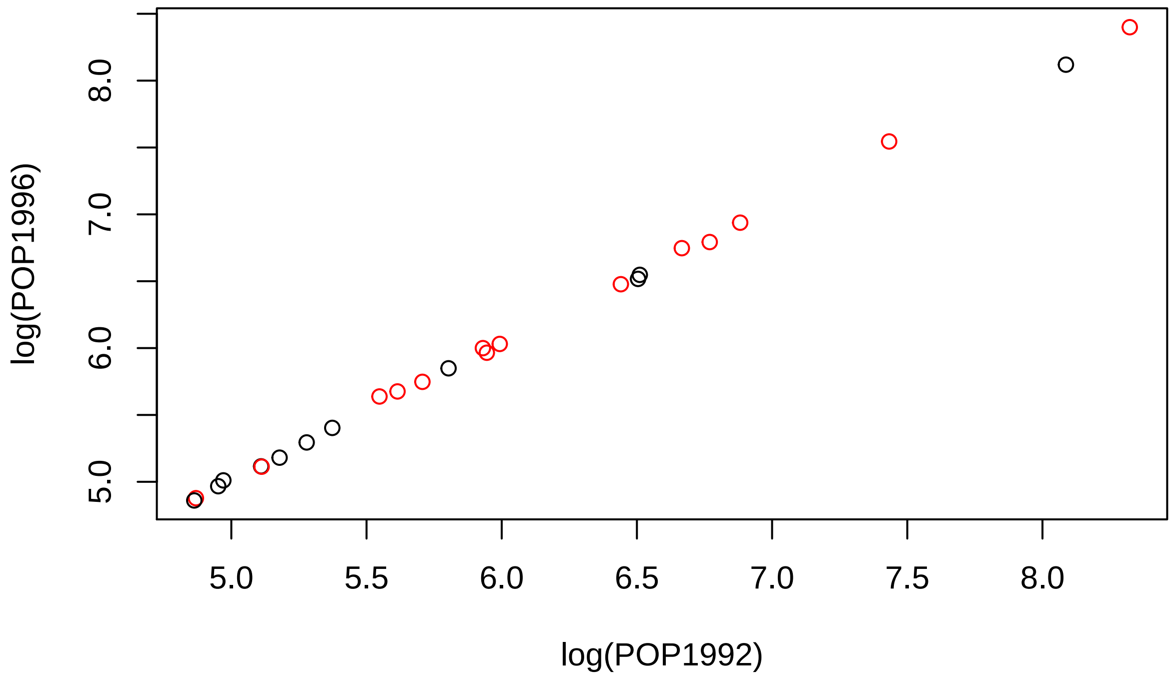
使用原始数据有

```
plot(POP1996 ~ POP1992, data = cities,
     col = as.integer(cities$have))
```



使用对数变换后的数据有

```
plot(log(POP1996) ~ log(POP1992), data = cities,
     col = as.integer(cities$have))
```

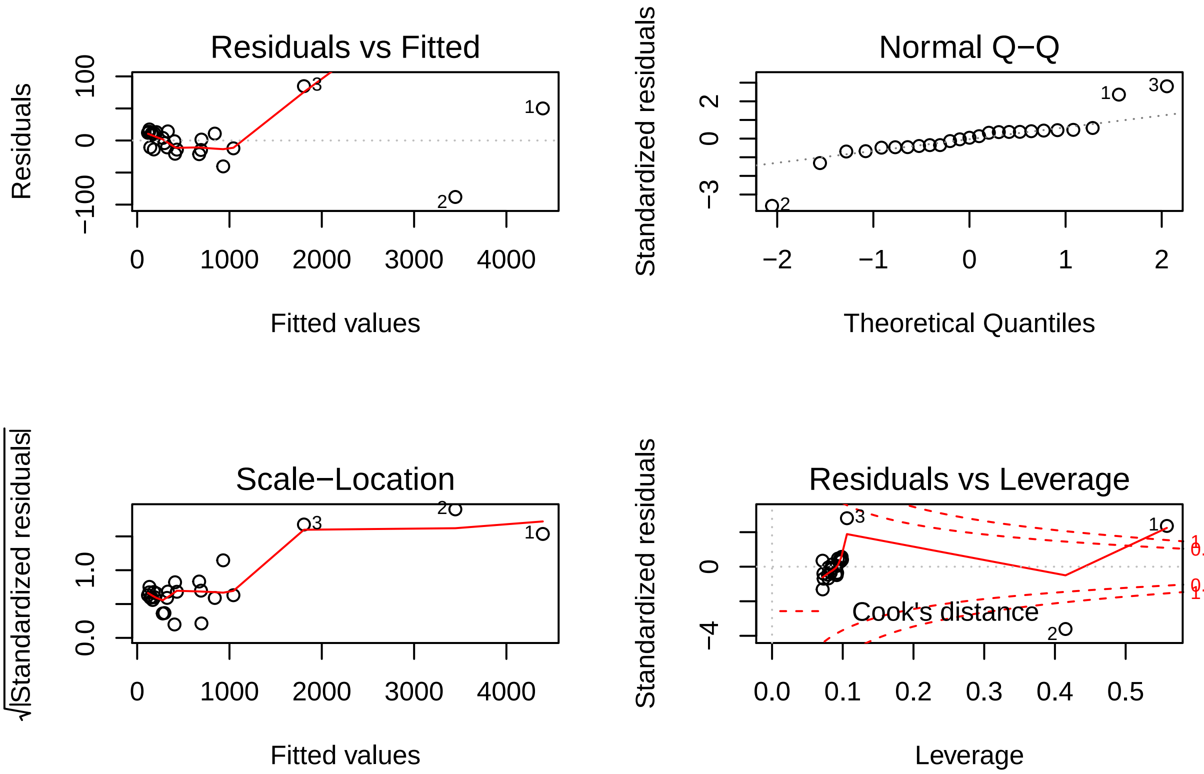


因为数据对数化后分布更加均匀，所以对数化后作图更好一点。

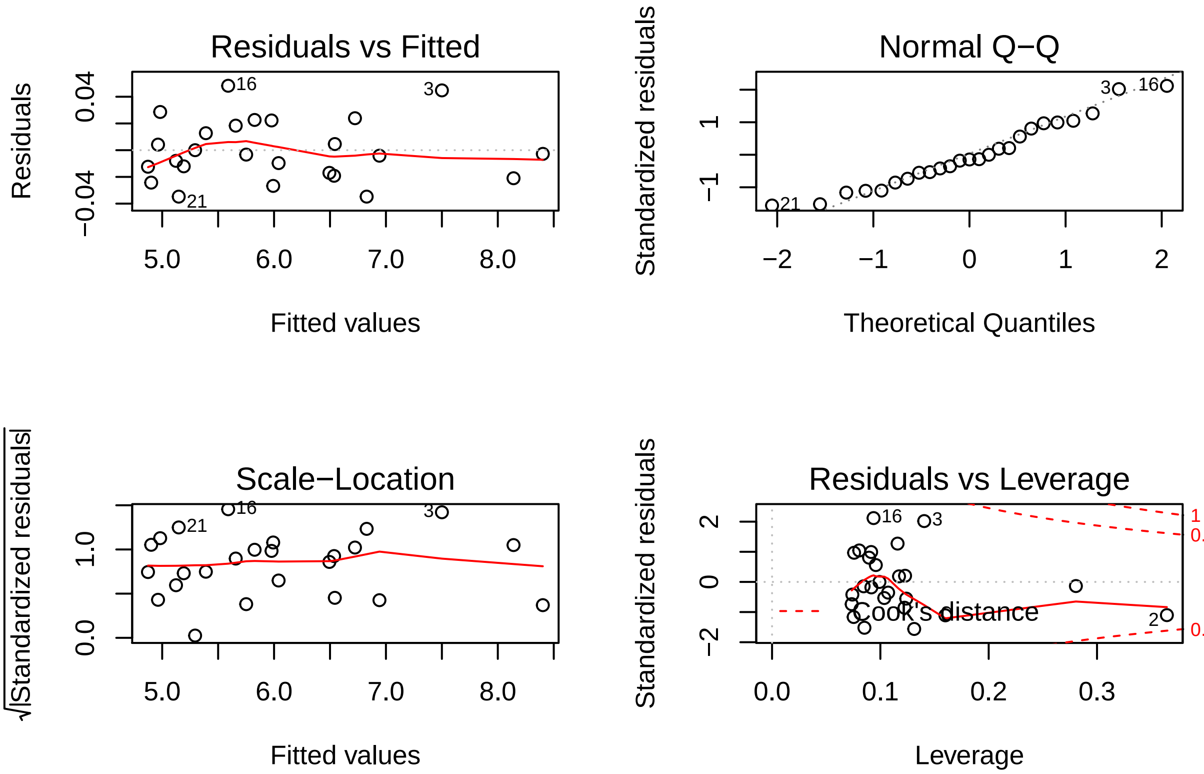
```
cities.lm1 <- lm(POP1996~have+POP1992, data = cities)
cities.lm2 <- lm(log(POP1996)~have+log(POP1992), data = cities)
```

两个回归的诊断图象如下所示，

```
par(mfrow=c(2,2))
plot(cities.lm1)
```




```
par(mfrow=c(2,2))
plot(cities.lm2)
```



比较这两个回归模型的诊断图可以看出，采用第二种模型会更好，首先，从残差图可以看

出模型 1 的残差远远高于模型 2，并且存在较多的异常值；另外，从 QQ 图可以看出模型 1 不满足残差正态性假设，与直线 $y = x$ 偏差较大，而相比于模型 1，模型 2 的 QQ 图基本上落在 $y = x$ 直线上。基于这两点，可以认定模型 2 优于模型 1。

2 MB 6.2

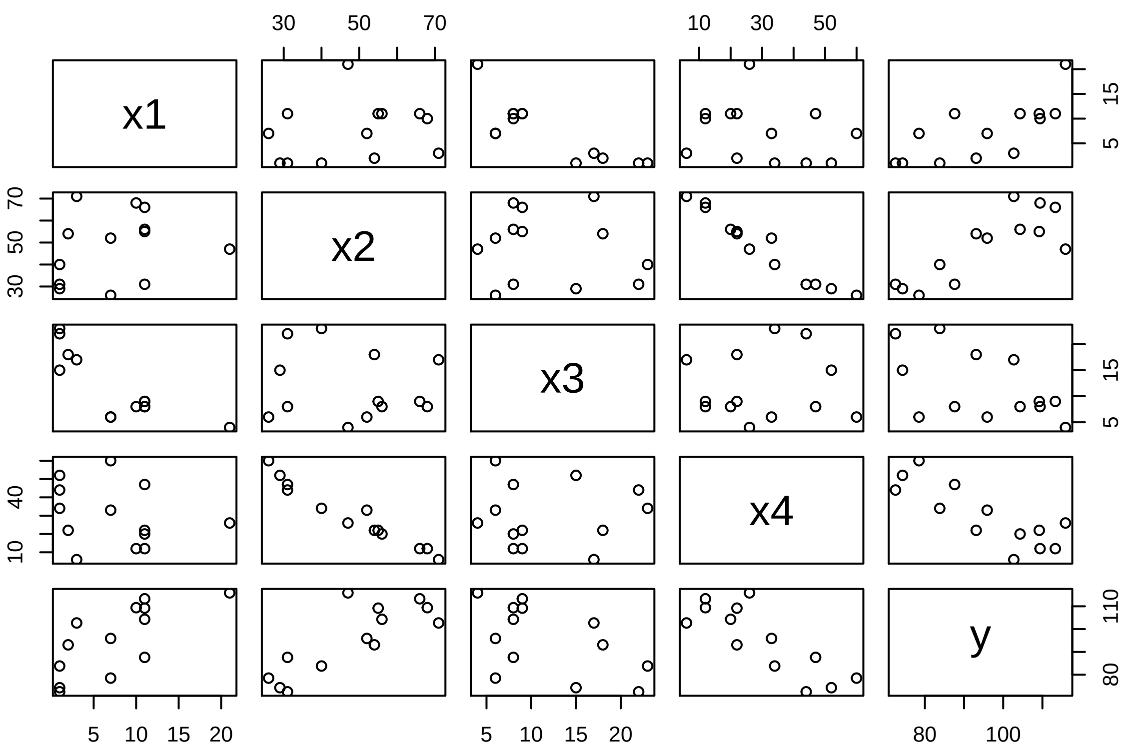
散点图矩阵如下

```r
library(MASS)
```

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

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

```r
pairs(cement)
```



从散点图矩阵中大致可以发现，y 与 x1，x2 正相关，而与 x3,x4 负相关。

进行多元回归有

```r
summary(lm(y~x1+x2+x3+x4, data = cement))
```

```
##
## Call:
```

```
## lm(formula = y ~ x1 + x2 + x3 + x4, data = cement)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1750 -1.6709  0.2508  1.3783  3.9254
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  62.4054    70.0710   0.891   0.3991
## x1            1.5511     0.7448   2.083   0.0708 .
## x2            0.5102     0.7238   0.705   0.5009
## x3            0.1019     0.7547   0.135   0.8959
## x4           -0.1441     0.7091  -0.203   0.8441
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.446 on 8 degrees of freedom
## Multiple R-squared:  0.9824, Adjusted R-squared:  0.9736
## F-statistic: 111.5 on 4 and 8 DF,  p-value: 4.756e-07
```

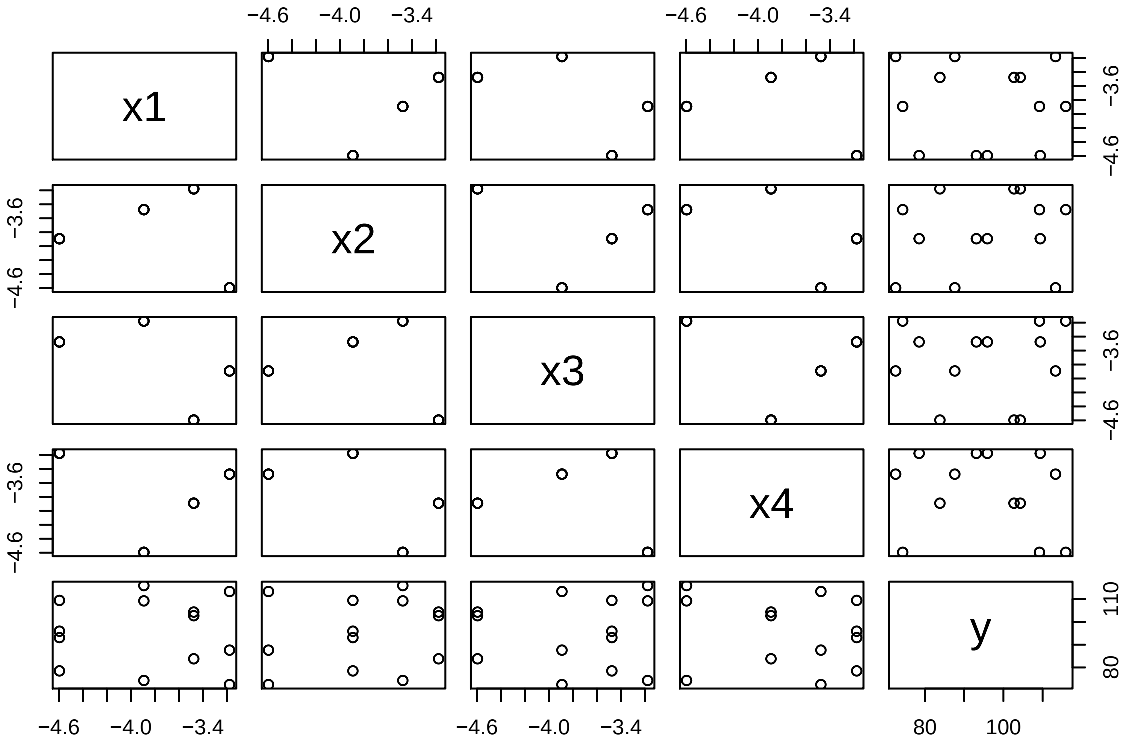
从回归结果可以看出，虽然 R^2 较高，但是各个系数的显著性水平都不高。

进行 $log(x/(100-x))$ 变换后

```
cement2 = cement
cement2[1:4] <- sapply(1:4, function(x) log(x/(100-x)))
```

此时散点图矩阵为

```
pairs(cement2)
```



进行多元回归我们有

```
summary(lm(y~x1+x2+x3+x4, data = cement))
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4, data = cement)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1750 -1.6709  0.2508  1.3783  3.9254
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  62.4054    70.0710   0.891   0.3991
## x1            1.5511     0.7448   2.083   0.0708 .
## x2            0.5102     0.7238   0.705   0.5009
## x3            0.1019     0.7547   0.135   0.8959
## x4           -0.1441     0.7091  -0.203   0.8441
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.446 on 8 degrees of freedom
## Multiple R-squared:  0.9824, Adjusted R-squared:  0.9736
## F-statistic: 111.5 on 4 and 8 DF,  p-value: 4.756e-07
```

结合回归结果和散点图矩阵来看，不进行变换的效果更好一点。

为了进一步研究，我们还可以考虑交叉项的影响，在构造多元回归的时候加入交叉项，可能会是模型更加完善。

3 MB 6.4

分别对男性女性的爬山时间进行回归分析，得到如下结果

```
lm.male = lm(time~dist+climb, data = hills2000)
lm.female = lm(timef~dist+climb, data = hills2000)
summary(lm.male)
```

```
##
## Call:
## lm(formula = time ~ dist + climb, data = hills2000)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.74979 -0.12722  0.01749  0.11139  0.69265
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.343e-01  4.694e-02  -7.121 2.88e-09 ***
## dist         1.655e-01  7.436e-03  22.264  < 2e-16 ***
## climb        4.446e-05  2.995e-05   1.485    0.144
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.213 on 53 degrees of freedom
## Multiple R-squared:  0.9673, Adjusted R-squared:  0.9661
## F-statistic: 785.1 on 2 and 53 DF,  p-value: < 2.2e-16
```

```
summary(lm.female)
```

```
##
## Call:
## lm(formula = timef ~ dist + climb, data = hills2000)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.76230 -0.32269 -0.01089  0.21152  1.73468
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) -6.519e-01  1.119e-01  -5.828 3.59e-07 ***
## dist          2.900e-01  1.755e-02  16.524  < 2e-16 ***
## climb        -1.168e-04  7.104e-05  -1.644    0.106
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5014 on 52 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.9273, Adjusted R-squared:  0.9245
## F-statistic: 331.7 on 2 and 52 DF,  p-value: < 2.2e-16
```

从回归结果中的 R^2 看，两个回归模型的拟合结果均较好。

4 MB 6.6

4.1 (a)

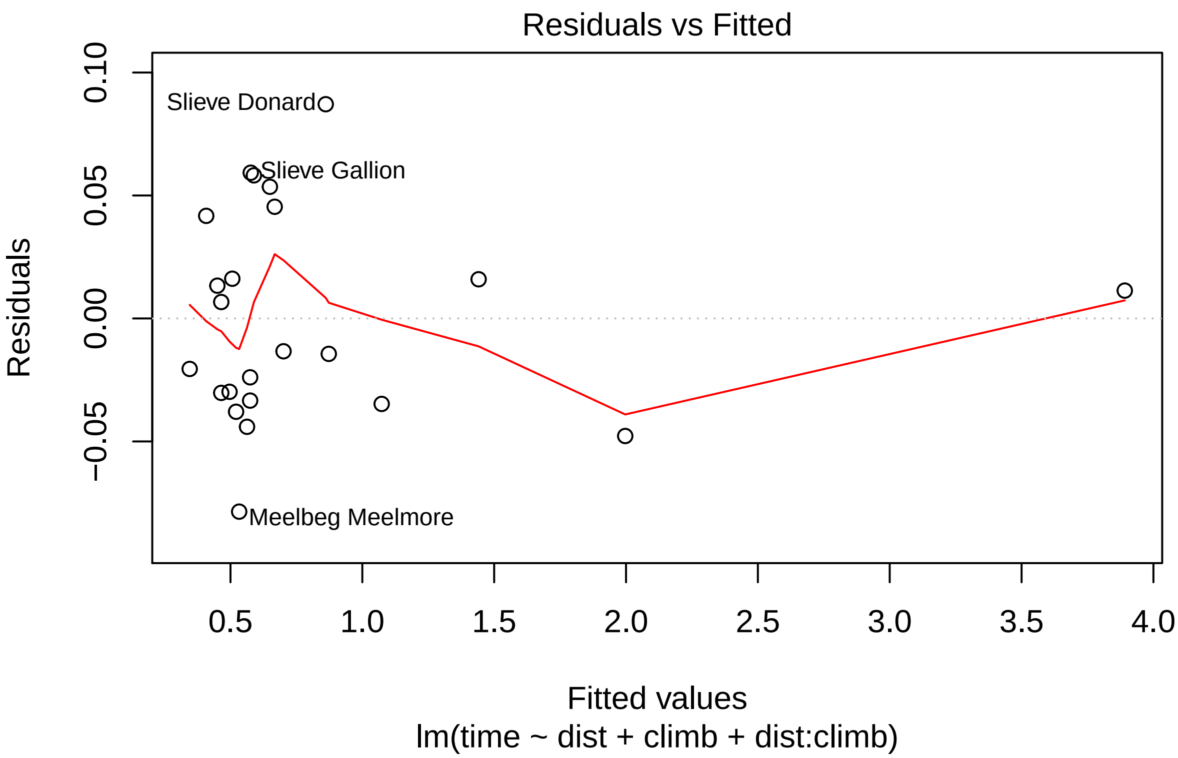
```r
nihills.lm <- lm(time ~ dist + climb, data = nihills)
nihills2.lm <- lm(time ~ dist + climb + dist:climb, data = nihills)
anova(nihills.lm, nihills2.lm)
```

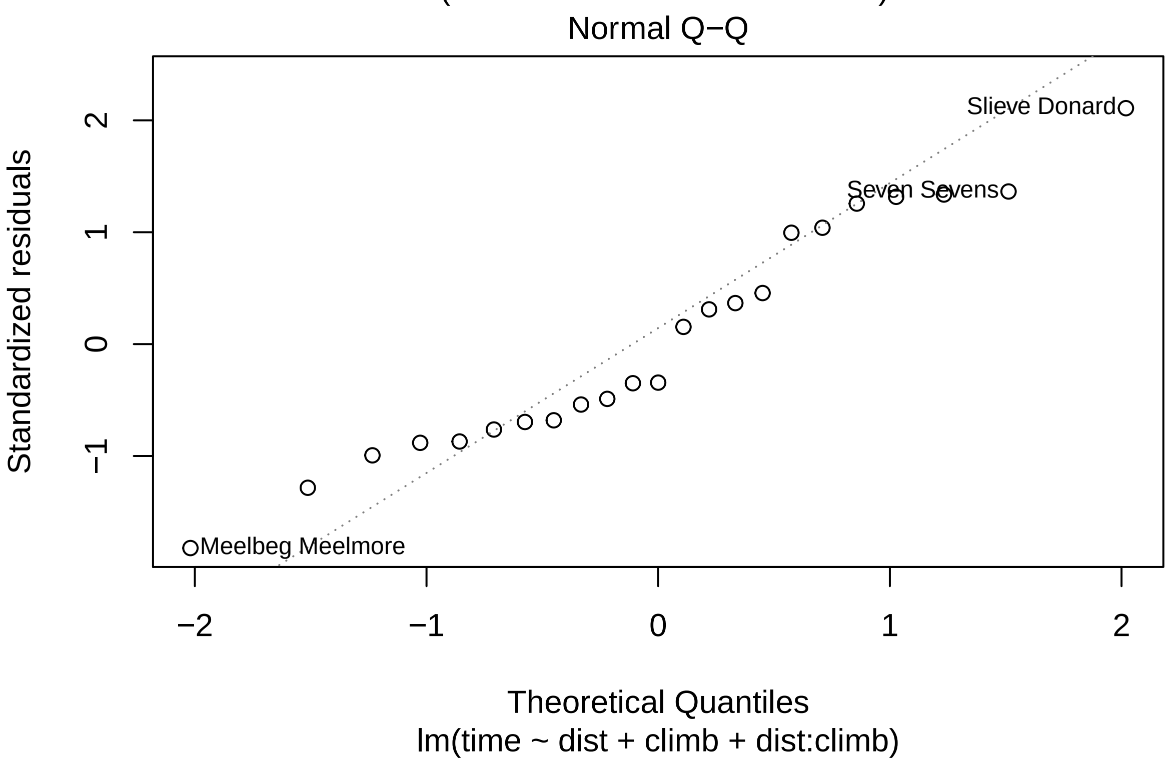
```
## Analysis of Variance Table
##
## Model 1: time ~ dist + climb
## Model 2: time ~ dist + climb + dist:climb
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1     20 0.189361
## 2     19 0.039361  1      0.15 72.406 6.623e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

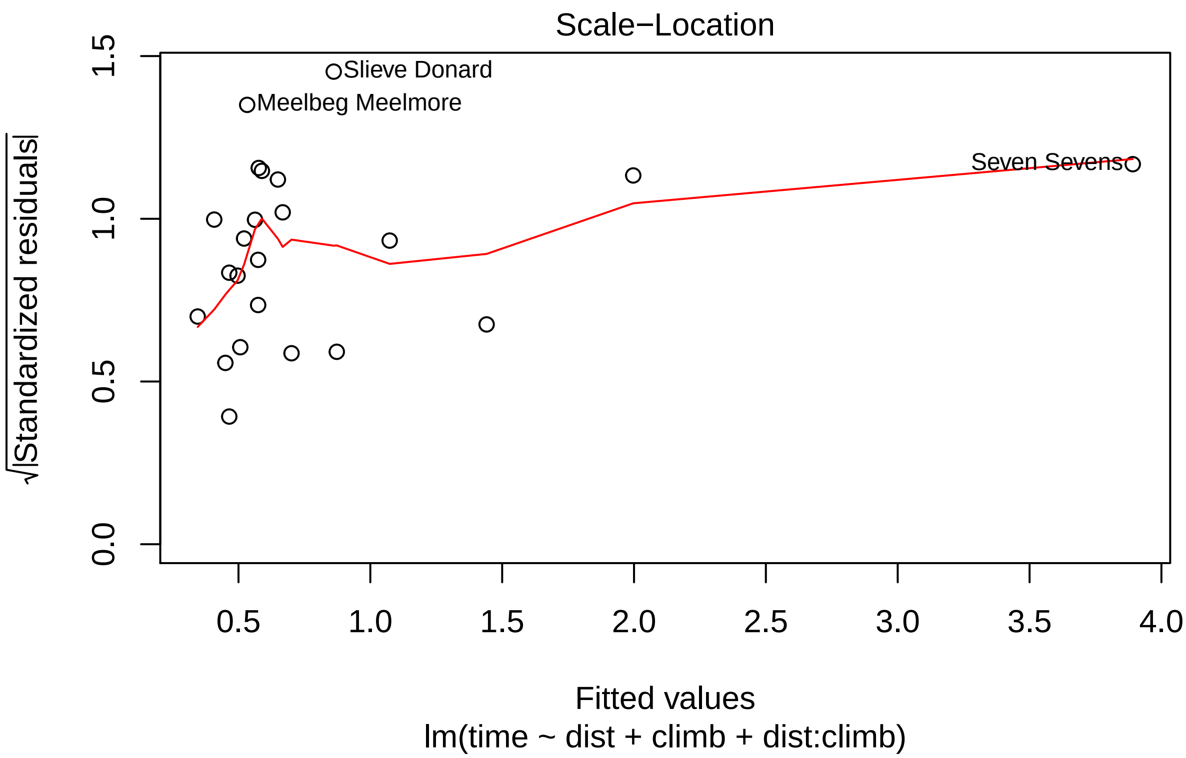
4.2 (b)

由 F 检验的结果只，模型 2 显著，于是选择模型 2。诊断图象如下

```r
plot(nihills2.lm)
```

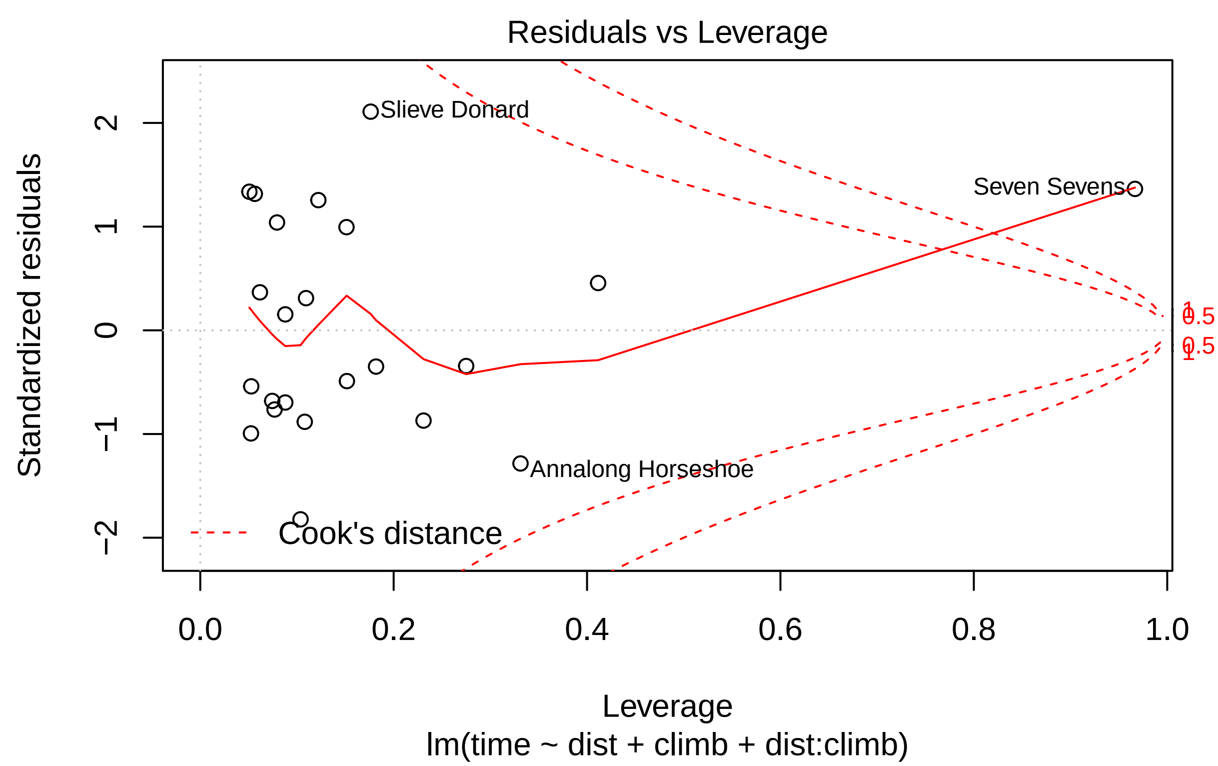

Residuals vs Fitted
Residuals
0.10
0.05
0.00
−0.05
Slieve Donard
Slieve Gallion
Meelbeg Meelmore
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Fitted values
lm(time ~ dist + climb + dist:climb)


Normal Q−Q
Standardized residuals
2
1
0
−1
Slieve Donard
Seven Sevens
Meelbeg Meelmore
−2
−1
0
1
2
Theoretical Quantiles
lm(time ~ dist + climb + dist:climb)




```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



从残差杠杆图可以看出 Seven Sevens 为异常点，因为其 cook 距离大于 1。

删掉该点

```r
nihills2 <- nihills[-which(rownames(nihills)=="Seven Sevens"),]
```

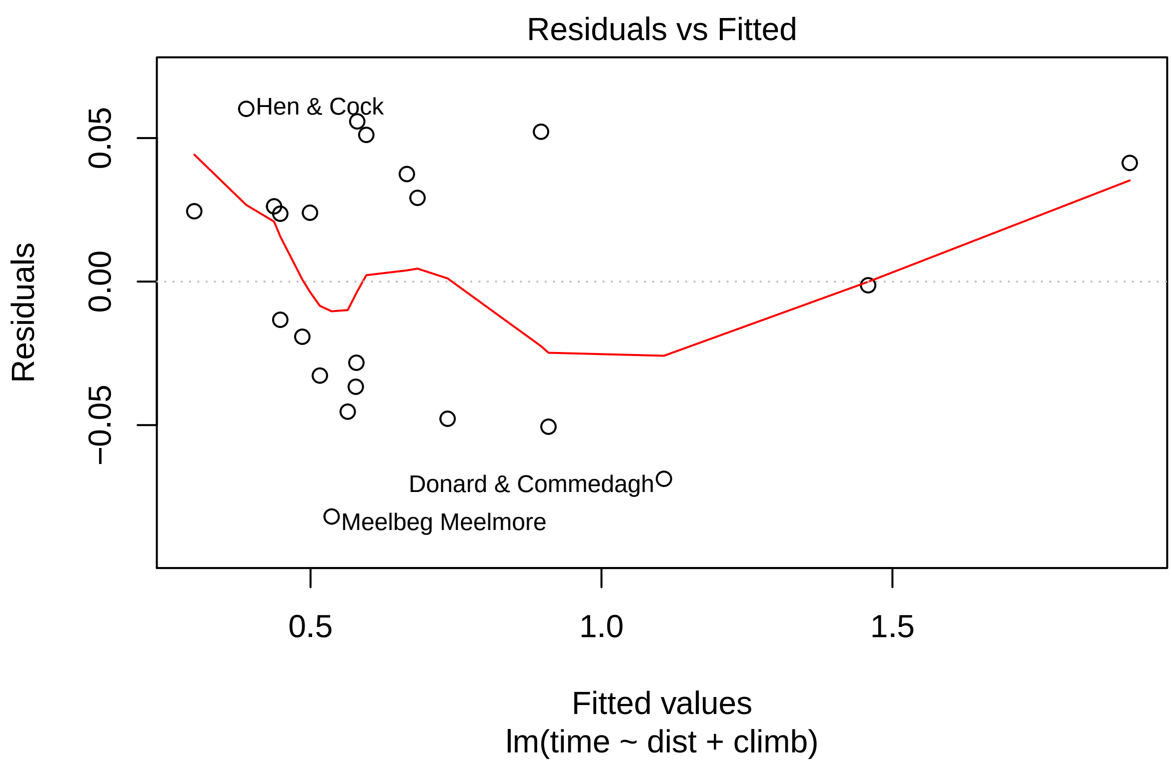
重新拟合模型

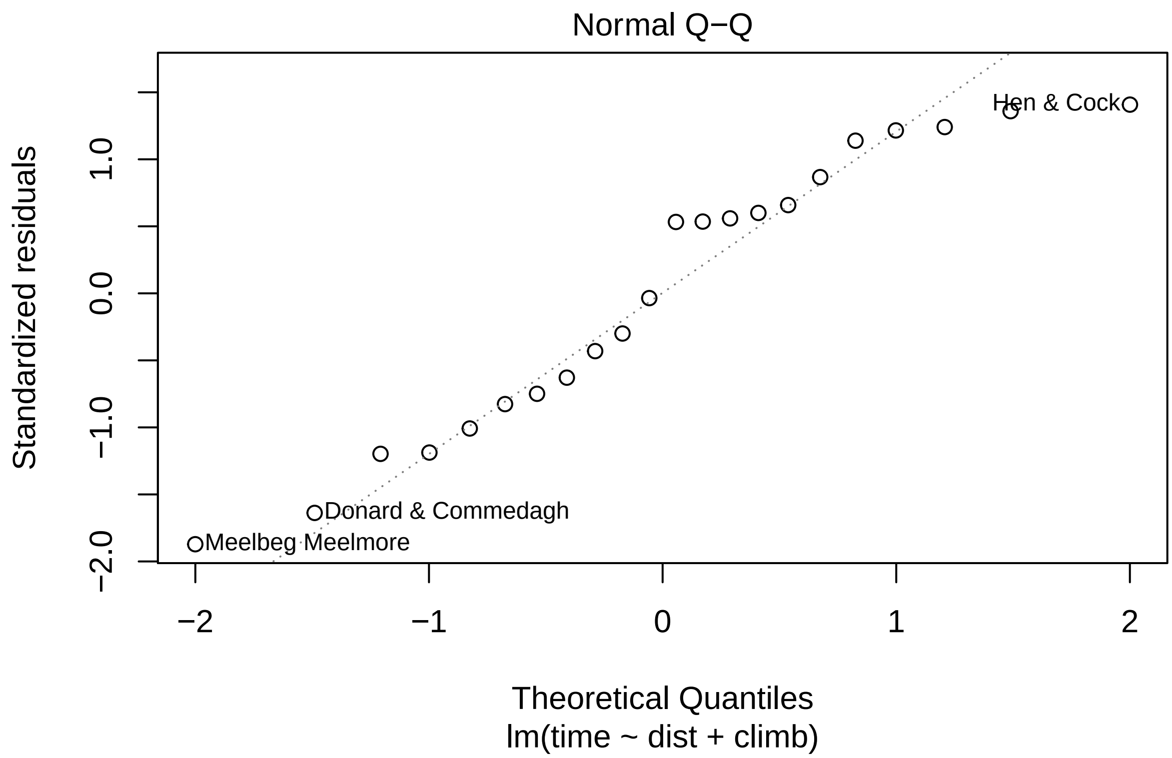
```r
nihills.lm.rm <- lm(time ~ dist + climb, data = nihills2)
nihills2.lm.rm <- lm(time ~ dist + climb + dist:climb, data = nihills2)
anova(nihills.lm.rm, nihills2.lm.rm)
```

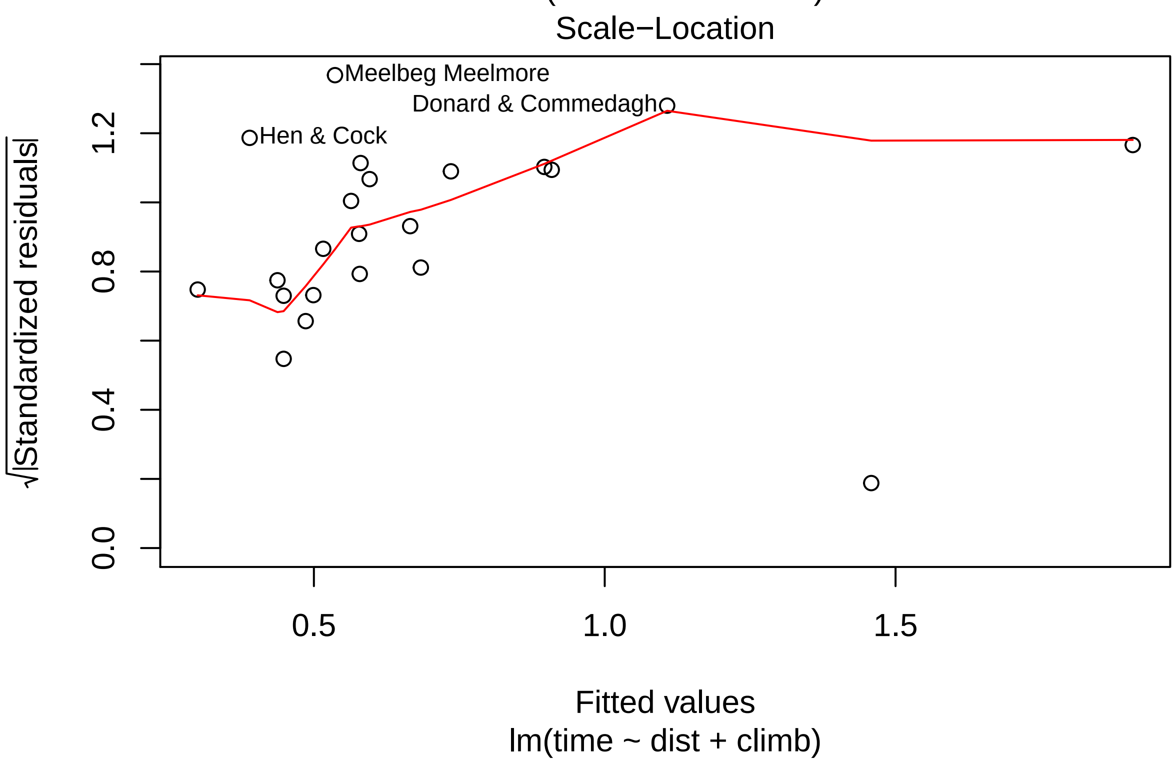
```
## Analysis of Variance Table
##
## Model 1: time ~ dist + climb
## Model 2: time ~ dist + climb + dist:climb
##   Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1     19 0.040533
## 2     18 0.035509  1 0.0050248 2.5472 0.1279
```

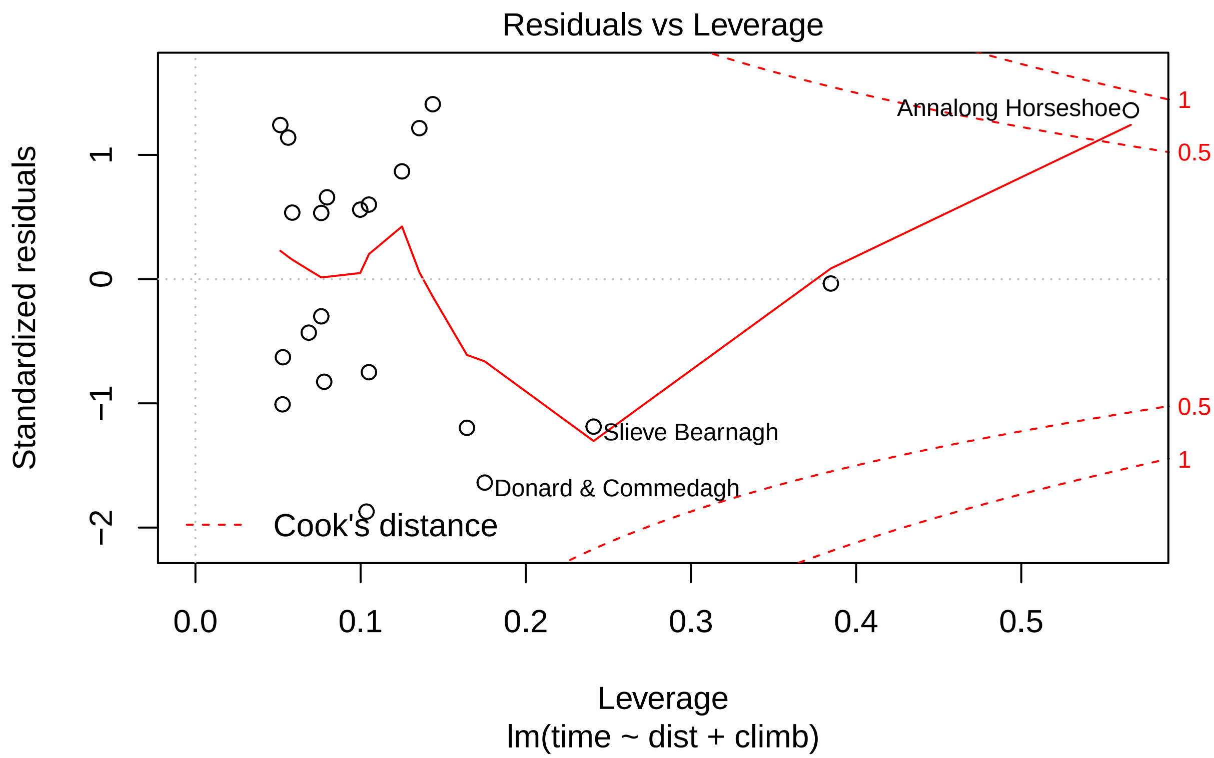
由 anova 分析结果，此时 p 值不显著，也就是交叉项不显著，因此此时应该采用模型 1，得到下面的诊断图：

```r
plot(nihills.lm.rm)
```




Normal Q−Q
Standardized residuals
−2.0
−1.0
0.0
1.0
−2
−1
0
1
2
Hen & Cock
Donard & Commedagh
Meelbeg Meelmore
Theoretical Quantiles
lm(time ~ dist + climb)


Scale−Location
√|Standardized residuals|
0.0
0.4
0.8
1.2
0.5
1.0
1.5
Meelbeg Meelmore
Donard & Commedagh
Hen & Cock
Fitted values
lm(time ~ dist + climb)



其中从残差杠杆图可以看出有一个点的 Cook 距离位于 0.5 和 1 之间，虽然相对偏大，但在容许 Cook 距离小于 1 的情形下可以不看成异常点。

5 MB 6.7

```
lm.litters <- lm(brainwt ~ bodywt + lsize, data = litters)
vif(lm.litters)
```

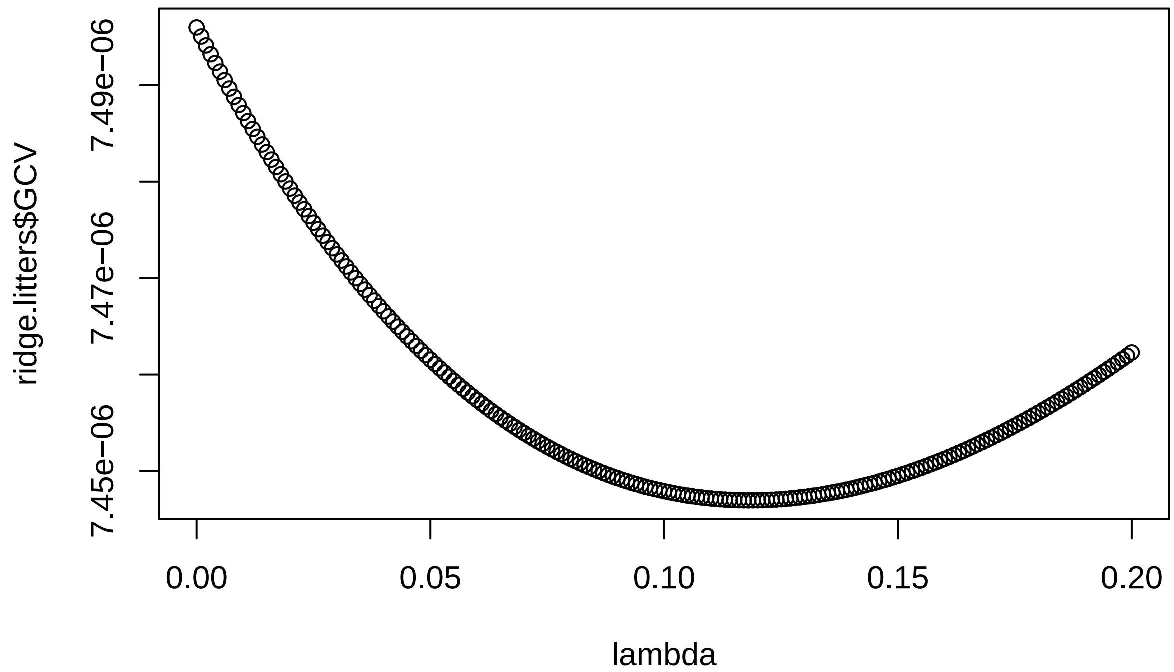
```
## bodywt  lsize
##  11.33  11.33
```

因 bodywy 和 lsize 的 VIF 都大于 10，则表明该模型有严重的多重共线性，于是需要进一步优化模型，如采用主成分回归。

6 MB 6.8

6.1 (a)

```
lambda = seq(0,0.2,0.001)
ridge.litters <- lm.ridge(brainwt ~ bodywt + lsize, data = litters, lambda = lambda)
plot(lambda, ridge.litters$GCV)
```



取 GCV 最低时的 λ 作为岭回归模型

```
lambda.min = lambda[which.min(ridge.litters$GCV)]
```

则此时岭回归模型为

```
ridge.litters.min = lm.ridge(brainwt ~ bodywt + lsize,
                             data = litters, lambda = lambda.min)
```

其变量系数为

```
coef(ridge.litters.min)
```

```
##                  bodywt       lsize
## 0.203442601 0.022050278 0.005661579
```

而 lm 的变量系数为

```
coef(lm.litters)
```

```
## (Intercept)      bodywt       lsize
## 0.178246962 0.024306344 0.006690331
```

可见，bodywt 的系数相差不大，但是 lsize 的系数岭回归更大。

6.2 (b)

```
## 岭回归估计
coef(ridge.litters.min) %*% c(1, 7, 10)
```

```
##             [,1]
## [1,] 0.4144103
```

```r
## 最小二乘回归估计
coef(lm.litters) %*% c(1, 7, 10)
```

```
##             [,1]
## [1,] 0.4152947
```

编写下面的 bootstrap.litter(B, seed) 函数，通过产生 B 个 bootstrap 样本，对每个 bootstrap 样本估计 mean brain weight 并返回。

```r
bootstrap.litter <- function(B, seed)
{
  set.seed(seed)
  Bsample = sapply(1:B, function(x) sample(nrow(litters), replace = TRUE))
  lambda = seq(0,0.2,0.001)
  res = c()
  for (i in 1:B)
  {
    lm.litters <- lm(brainwt ~ bodywt + lsize, data = litters[Bsample[,i], ])
    ridge.litters.min = lm.ridge(brainwt ~ bodywt + lsize,
                                 data = litters[Bsample[,i], ], lambda = lambda.min)
    lambda.min = lambda[which.min(ridge.litters$GCV)]
    ridge.litters.min = lm.ridge(brainwt ~ bodywt + lsize,
                                 data = litters[Bsample[,i], ], lambda = lambda.min)
    res = rbind(res, c(coef(lm.litters) %*% c(1, 7, 10),
                       coef(ridge.litters.min) %*% c(1, 7, 10)))
  }
  return(res)
}
```

下面求 B 个 bootstrap 样本的 0.025 和 0.975 分位数，从而得到 95% 的置信区间。

```r
bootstrap.res = bootstrap.litter(1000, 123)
q.res = apply(bootstrap.res, 2, function(x) quantile(x, c(0.025, 0.975)))
q.res
```

```
##            [,1]      [,2]
## 2.5%  0.4058261 0.4043712
## 97.5% 0.4231659 0.4222348
```

于是通过 bootstrap 求得的最小二乘回归的 95% 置信区间为 $[0.4058261, 0.4231659]$，岭回归的 95% 置信区间为 $[0.4043712, 0.4222348]$。而通过 predict.lm 求得的 95% 置信区间为 $[0.4062582, 0.4243312]$，

```r
predict.lm(lm.litters, data.frame(lsize=10, bodywt = 7), interval = "confidence")
```

```
##         fit       lwr       upr
```

```
## 1 0.4152947 0.4062582 0.4243312
```

值得说明的是，predict.lm 没有针对 ridgelm 的方法，故无法计算，只能比较最小二乘估计的 bootstrap 方法和 predict.lm 方法的置信区间。

7 MDL Chapter 14 Worksheet B: Study of intima media

下载数据

```
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)
```

注意到每次 tobacco 取值为 0（表明为非吸烟者）时，则 packyear 值为 NA（表示每年香烟的盒数），所以很自然地可以将 packyear 的 NA 换成 0。

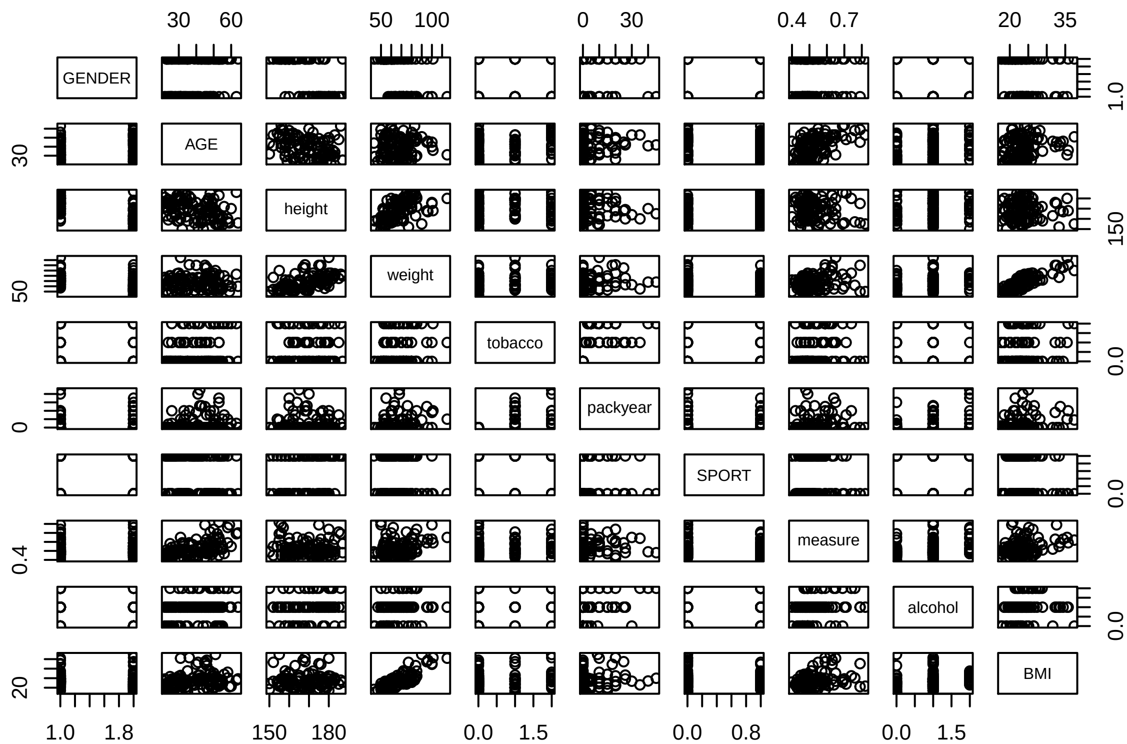
```
data$packyear[is.na(data$packyear)] <-0
```

并且计算 BMI

```
data = within(data, {
  BMI = weight/(height/100)^2
})
```

7.1 Problem 14.1

```
pairs(data)
```



从散点图矩阵可以看出，height 和 weight 可能存在多重共线性，BMI 和 weight 可能存在多重共线性，因为它们两两间有较大的线性关系。

7.2 Problem 14.2

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

```
##
## 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
```

measure 和 SPORT 的回归模型为

```r
lm.sport <- lm(measure~SPORT, data)
summary(lm.sport)
```

```
##
## Call:
## lm(formula = measure ~ SPORT, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.12082 -0.05787 -0.02434  0.04918  0.28213
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.53787    0.01109  48.490   <2e-16 ***
## SPORT       -0.01705    0.01662  -1.026    0.307
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08663 on 108 degrees of freedom
## Multiple R-squared:  0.009654,   Adjusted R-squared:  0.0004839
## F-statistic: 1.053 on 1 and 108 DF,  p-value: 0.3072
```

measure 和 alcohol 的回归模型为

```r
lm.alcohol <- lm(measure~alcohol, data)
summary(lm.alcohol)
```

```
##
## Call:
## lm(formula = measure ~ alcohol, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.13245 -0.05567 -0.02031  0.03504  0.28754
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.49817    0.01510  33.000   <2e-16 ***
## alcohol      0.03429    0.01363   2.516   0.0133 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.08461 on 108 degrees of freedom
## Multiple R-squared:  0.05537,    Adjusted R-squared:  0.04663
## F-statistic: 6.331 on 1 and 108 DF,  p-value: 0.01333
```

measure 和 packyear 的回归模型为

```r
lm.packyear <- lm(measure~packyear, data)
summary(lm.packyear)
```

```
##
## Call:
## lm(formula = measure ~ packyear, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.13202 -0.05337 -0.02480  0.03764  0.29663
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.5233716  0.0092582  56.531   <2e-16 ***
## packyear    0.0014323  0.0008909   1.608    0.111
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08603 on 108 degrees of freedom
## Multiple R-squared:  0.02337,    Adjusted R-squared:  0.01433
## F-statistic: 2.585 on 1 and 108 DF,  p-value: 0.1108
```

measure 和 BMI 的回归模型为

```r
lm.BMI <- lm(measure~BMI, data)
summary(lm.BMI)
```

```
##
## Call:
## lm(formula = measure ~ BMI, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.12510 -0.05842 -0.02021  0.03050  0.30953
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.370113   0.048395   7.648 9.01e-12 ***
## BMI         0.006744   0.002010   3.354   0.0011 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.08285 on 108 degrees of freedom
## Multiple R-squared:  0.09436,    Adjusted R-squared:  0.08597
## F-statistic: 11.25 on 1 and 108 DF,  p-value: 0.001098
```

7.3 Problem 14.3

上问中 p<0.25 的变量有 AGE、alcohol、packyear 和 BMI

```
lm.age.packyear <- lm(measure~AGE*packyear, data)
summary(lm.age.packyear)
```

```
##
## Call:
## lm(formula = measure ~ AGE * packyear, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.15508 -0.05035 -0.00294  0.03038  0.23904
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.3654531  0.0281989  12.960  < 2e-16 ***
## AGE           0.0041223  0.0007043   5.853 5.45e-08 ***
## packyear     -0.0010219  0.0045630  -0.224    0.823
## AGE:packyear  0.0000338  0.0001038   0.326    0.745
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0731 on 106 degrees of freedom
## Multiple R-squared:  0.308,  Adjusted R-squared:  0.2884
## F-statistic: 15.72 on 3 and 106 DF,  p-value: 1.577e-08
```

```
lm.alcohol.packyear <- lm(measure~alcohol*packyear, data)
summary(lm.alcohol.packyear)
```

```
##
## Call:
## lm(formula = measure ~ alcohol * packyear, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.14302 -0.05474 -0.01615  0.03425  0.29304
##
## Coefficients:
```

```
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.485350   0.016211  29.940  < 2e-16 ***
## alcohol           0.041604   0.015114   2.753  0.00695 **
## packyear          0.004278   0.001991   2.149  0.03394 *
## alcohol:packyear -0.002475   0.001313  -1.886  0.06205 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0836 on 106 degrees of freedom
## Multiple R-squared:  0.09484,    Adjusted R-squared:  0.06922
## F-statistic: 3.702 on 3 and 106 DF,  p-value: 0.01402
```

```r
lm.BMI.packyear <- lm(measure~BMI*packyear, data)
summary(lm.BMI.packyear)
```

```
##
## Call:
## lm(formula = measure ~ BMI * packyear, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.12046 -0.05558 -0.01667  0.02698  0.31551
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.3513526  0.0622731   5.642 1.41e-07 ***
## BMI           0.0073582  0.0026388   2.788  0.00628 **
## packyear      0.0045305  0.0059781   0.758  0.45022
## BMI:packyear -0.0001462  0.0002422  -0.604  0.54725
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08301 on 106 degrees of freedom
## Multiple R-squared:  0.1076, Adjusted R-squared:  0.08234
## F-statistic:  4.26 on 3 and 106 DF,  p-value: 0.006975
```

7.4 Problem 14.4

25% 显著的单变量有 AGE、alcohol、packyear 和 BMI，10% 显著的交叉变量有 alcohol*packyear

则模型为

```r
lm.all <- lm(measure~AGE+alcohol*packyear+BMI, data)
summary(lm.all)
```

```
##
## Call:
## lm(formula = measure ~ AGE + alcohol * packyear + BMI, data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.152054 -0.039628 -0.009446  0.032913  0.250509
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.2625953  0.0451210   5.820 6.59e-08 ***
## AGE               0.0037861  0.0006383   5.931 3.97e-08 ***
## alcohol           0.0192951  0.0132043   1.461   0.1470
## packyear          0.0020499  0.0017268   1.187   0.2379
## BMI               0.0041774  0.0017839   2.342   0.0211 *
## alcohol:packyear -0.0014955  0.0011237  -1.331   0.1862
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07059 on 104 degrees of freedom
## Multiple R-squared:  0.3669, Adjusted R-squared:  0.3364
## F-statistic: 12.05 on 5 and 104 DF,  p-value: 3.245e-09
```

7.5 Problem 14.5

由上述 summary 的结果可以看出，此时交叉项不再显著，则除掉交叉项为

```
lm.all2 <- lm(measure~AGE+alcohol+packyear+BMI, data)
summary(lm.all2)
```

```
##
## Call:
## lm(formula = measure ~ AGE + alcohol + packyear + BMI, data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.152784 -0.045663 -0.008005  0.035597  0.250100
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.601e-01  4.525e-02   5.748 8.93e-08 ***
## AGE          3.844e-03  6.391e-04   6.015 2.66e-08 ***
## alcohol      1.215e-02  1.211e-02   1.004   0.3179
## packyear    -6.959e-06  7.729e-04  -0.009   0.9928
```

```
## BMI          4.513e-03  1.772e-03   2.546   0.0123 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07085 on 105 degrees of freedom
## Multiple R-squared:  0.3561, Adjusted R-squared:  0.3316
## F-statistic: 14.52 on 4 and 105 DF,  p-value: 1.81e-09
```

7.6 Problem 14.6

由上述 summary 结果知 alcohol 和 packyear 不再显著，删掉 alcohol 有

```r
lm.all2.rm.alcohol <- lm(measure~AGE+packyear+BMI, data)
summary(lm.all2.rm.alcohol)
```

```
##
## Call:
## lm(formula = measure ~ AGE + packyear + BMI, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.15193 -0.04832 -0.00529  0.03567  0.25041
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.2623458  0.0451913   5.805 6.77e-08 ***
## AGE         0.0039677  0.0006272   6.325 6.14e-09 ***
## packyear    0.0001549  0.0007559   0.205  0.83800
## BMI         0.0046588  0.0017665   2.637  0.00962 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07085 on 106 degrees of freedom
## Multiple R-squared:  0.3499, Adjusted R-squared:  0.3315
## F-statistic: 19.02 on 3 and 106 DF,  p-value: 6.087e-10
```

虽然此时 packyear 也不显著，但题目要求不能改变与 tobacco 有关的变量，及不能改变 packyear 的变量，故保留。

7.7 Problem 14.7

最终模型是

$$\text{measure} \sim \text{AGE} + \text{BMI} + \text{packyear}$$

并且注意到，无论 packyear 的值为多少，measure 总会随着 AGE 和 BMI 的增大而增大。