Regression Lab 2

PB19151769 马宇骁

例 2.2/4.5 城镇人均收支分析

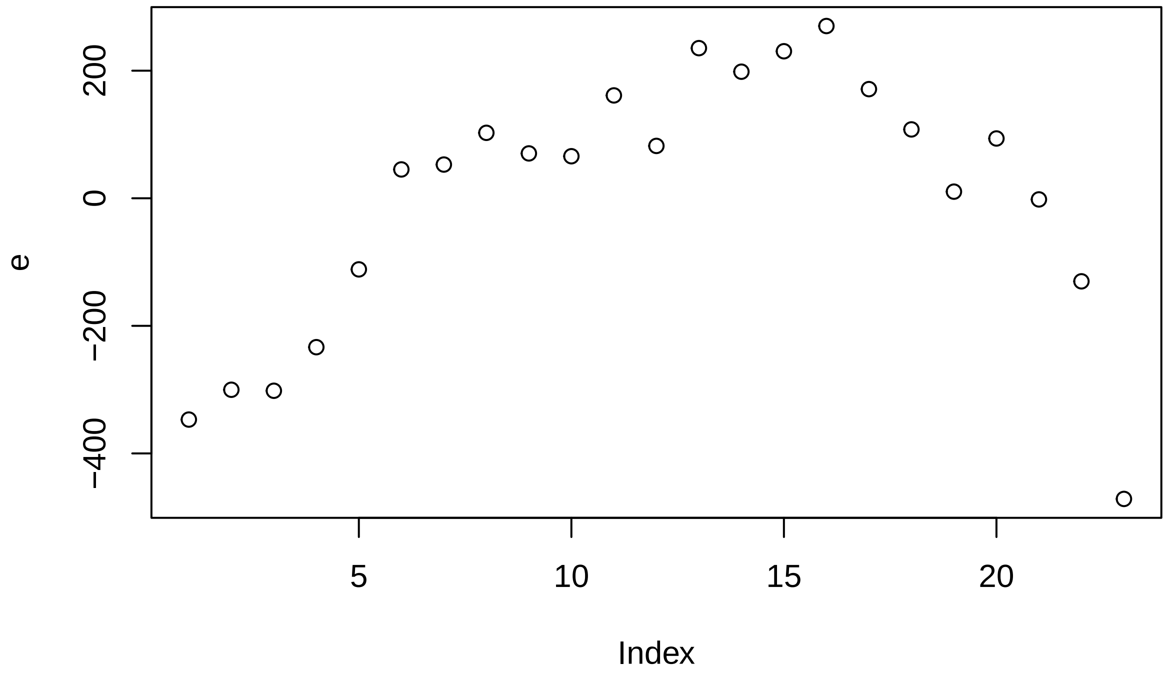
1. 用最小二次乘法做一元线性回归

```
y <- c(1278.89, 1453.8, 1671.7, 2110.8, 2851.3, 3537.57, 3919.5, 4185.6, 4331.6, 4615.9, 4998, 530
x <- c(1510.16, 1700.6, 2026.6, 2577.4, 3496.2, 4282.95, 4838.9, 5160.3, 5425.1, 5854, 6279.98, 68

olsy <- lm(y~x)
print(summary(olsy))
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -471.35 -120.86   65.89  134.58  269.99
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.092e+02  7.584e+01   8.033 7.71e-08 ***
## x           6.732e-01  6.762e-03  99.554  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 211.1 on 21 degrees of freedom
## Multiple R-squared:  0.9979, Adjusted R-squared:  0.9978
## F-statistic:  9911 on 1 and 21 DF,  p-value: < 2.2e-16
```

```
e<-summary(olsy)$resid # 提取残差序列
plot(e)
```



```
# print(e)
```

2. 计算 DW 值和 值，并判断误差项是否存在自相关

```
library(lmtest)
```

```
## Warning: 程辑包'lmtest'是用R版本4.1.3 来建造的
```

```
## 载入需要的程辑包：zoo
```

```
##
## 载入程辑包：'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
dw <- dwtest(olsy)
# print(dw)
DW <- (dw$statistic)
rho <- 1 - 0.5*DW
names(rho) = c('rho')
print(c(DW,rho))
```

```
##        DW       rho
## 0.2831283 0.8584358
```

```
dw
```

```
##
##  Durbin-Watson test
##
## data:  olsy
## DW = 0.28313, p-value = 7.712e-10
## alternative hypothesis: true autocorrelation is greater than 0
```

看到 p-value 远小于 0.05，故存在一阶自相关，且为正自相关。

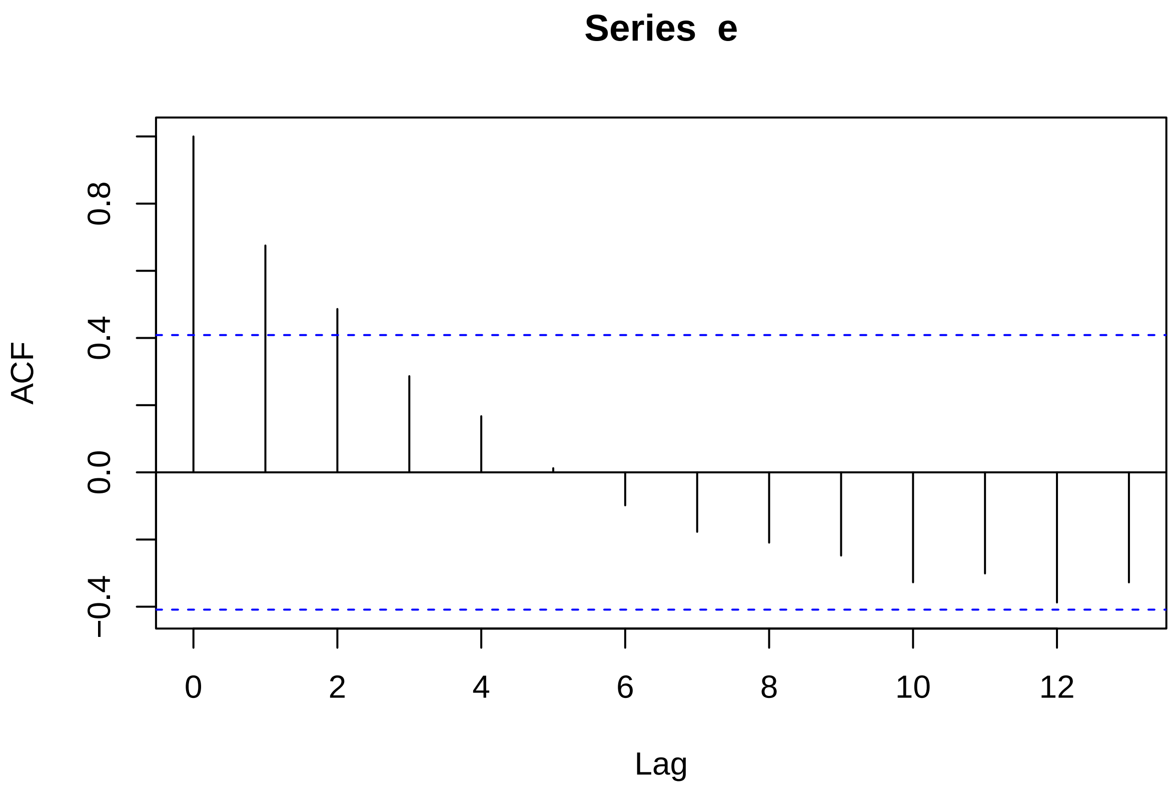
3. 用迭代法消除自相关

```
library(car)
```

```
## Warning: 程辑包'car'是用R版本4.1.3 来建造的
```

```
## 载入需要的程辑包：carData
```

```
acf.1<-acf(e) # 计算各阶自相关系数;acf.1;acf.1$acf[[2]]
```



```
rhohat <- 1-0.2831283/2;rhohat
```

```
## [1] 0.8584359
```

```
newy<-y[2:length(y)]-rhohat*y[1:length(y)-1]
newx<-x[2:length(y)]-rhohat*x[1:length(y)-1]
new.reg<-lm(newy~newx)
summary(new.reg)
```

```
##
## Call:
## lm(formula = newy ~ newx)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -193.540  -38.320    0.731   55.922  150.349
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 185.34316   32.54827   5.694 1.42e-05 ***
## newx          0.62780    0.01198  52.421  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 86.31 on 20 degrees of freedom
## Multiple R-squared:  0.9928, Adjusted R-squared:  0.9924
## F-statistic:  2748 on 1 and 20 DF,  p-value: < 2.2e-16
```

```
durbinWatsonTest(new.reg)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1     -0.05907877      1.820341   0.512
##  Alternative hypothesis: rho != 0
```

看到，迭代一次之后的 DW 检验的 p-value 就已经远远大于 0.05，因此此时不能在 95% 的置信水平拒绝原假设，可以认为此时迭代已结束，序列自相关已消除。

4. 用差分法消除自相关

```
diffy <- diff(y)
diffx <- diff(x)
diff.reg <- lm(diffy~diffx)
summary(diff.reg)
```

```
##
## Call:
## lm(formula = diffy ~ diffx)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -170.36  -58.53   -2.22   39.91  139.14
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 95.70082   31.25005   3.062  0.00615 **
## diffx        0.57646    0.02419  23.826 3.73e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 85.69 on 20 degrees of freedom
## Multiple R-squared:  0.966,  Adjusted R-squared:  0.9643
## F-statistic: 567.7 on 1 and 20 DF,  p-value: 3.729e-16
```

```
durbinWatsonTest(diff.reg)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      -0.1559281      2.107853   0.982
##  Alternative hypothesis: rho != 0
```

看到，一次差分之后的 DW 检验的 p-value 就已经远远大于 0.05，因此此时不能在 95% 的置信水平拒绝原假设，可以认为序列自相关已消除。

5. 利用前两位所得方程做预测

根据第 3、4 题求出的回归方程，可以看出：在迭代法中

$$\hat{y_t} = 0.62780(x_t - 0.8584358 * x_{t-1}) + 0.8584358 * y_{t-1} + 185.34316$$

在差分法中

$$\hat{y_t} = 0.57646(x_t - x_{t-1}) + y_{t-1} + 95.70082$$

例 5.2/3.1 城镇居民消费因素分析

```
x1 <- c(7535,7344,4211,3856,5463,5809,4635,4687,9656,6658,7552,5815,7317,5072,5201,4607,5838,5442,
x2 <- c(2639,1881,1542,1529,2730,2042,2045,1807,2111,1916,2110,1541,1634,1477,2197,1886,1783,1625,
x3 <- c(1971,1854,1502,1439,1584,1433,1594,1337,1790,1437,1552,1397,1754,1174,1572,1191,1371,1302,
x4 <- c(1658,1556,1047,906,1354,1310,1448,1181,1017,1058,1228,1143,773,671,1005,1085,1030,918,1048
x5 <- c(3696,2254,1204,1506,1972,1844,1643,1217,3724,3078,2997,1933,2105,1487,1656,1525,1652,1738,
x6 <- c(84742,61514,38658,44236,46557,41858,38407,36406,78673,50639,50197,44601,44525,38512,41904,
x7 <- c(87475,93173,36584,33628,63886,56649,43415,35711,85373,68347,63374,28792,52763,28800,51768,
x8 <- c(106.5,107.5,104.1,108.8,109.6,107.7,111,104.8,106,112.6,104.5,105.3,104.6,106.7,106.9,106.
x9 <- c(1.3,3.6,3.7,3.3,3.7,3.6,3.7,4.2,3.1,3.1,3,3.7,3.6,3,3.3,3.1,3.8,4.2,2.5,3.4,2,3.3,4,3.3,4,
y <- c(24046,20024,12531,12212,17717,16594,14614,12984,26253,18825,21545,15012,18593,12776,15778,1
```

1. 计算相关系数矩阵

```
df <- data.frame(y,x1,x2,x3,x4,x5,x6,x7,x8,x9)
cor_ <- cor(df, method = 'pearson')
cor_
```

```
##             y         x1         x2         x3          x4         x5
## y   1.0000000  0.9022762 0.51172104  0.7811370  0.49423568  0.9414255
## x1  0.9022762  1.0000000 0.22714089  0.6117634  0.21301742  0.7872537
## x2  0.5117210  0.2271409 1.00000000  0.3053681  0.64622334  0.4704869
## x3  0.7811370  0.6117634 0.30536809  1.0000000  0.58409947  0.7364894
## x4  0.4942357  0.2130174 0.64622334  0.5840995  1.00000000  0.4881049
## x5  0.9414255  0.7872537 0.47048687  0.7364894  0.48810487  1.0000000
## x6  0.7848767  0.6967609 0.46044177  0.5392700  0.38109255  0.7468939
## x7  0.8733947  0.6970034 0.61465733  0.7768628  0.65131022  0.8141689
## x8 -0.1302697 -0.1633994 0.14367061 -0.1783940  0.07004622 -0.1043261
## x9 -0.3614779 -0.3755017 0.01334004 -0.3247017 -0.10969051 -0.3743180
##             x6         x7          x8          x9
## y   0.78487674  0.8733947 -0.13026967 -0.36147795
## x1  0.69676095  0.6970034 -0.16339935 -0.37550174
## x2  0.46044177  0.6146573  0.14367061  0.01334004
## x3  0.53927001  0.7768628 -0.17839396 -0.32470168
## x4  0.38109255  0.6513102  0.07004622 -0.10969051
## x5  0.74689394  0.8141689 -0.10432612 -0.37431801
## x6  1.00000000  0.7801488 -0.01790576 -0.49913442
## x7  0.78014879  1.0000000 -0.01989850 -0.26236608
## x8 -0.01790576 -0.0198985  1.00000000 -0.13009092
## x9 -0.49913442 -0.2623661 -0.13009092  1.00000000
```

2. 对全模型做线性回归拟合

```
olsa <- lm(y~., data=df)
summary((olsa))
```

```
##
## Call:
## lm(formula = y ~ ., data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -940.13 -195.24    3.42  239.00  476.06
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.206e+02  3.952e+03   0.081 0.936097
## x1           1.317e+00  1.062e-01  12.400 3.97e-11 ***
## x2           1.650e+00  3.008e-01   5.484 1.93e-05 ***
## x3           2.179e+00  5.199e-01   4.190 0.000412 ***
## x4          -5.609e-03  4.766e-01  -0.012 0.990720
## x5           1.684e+00  2.142e-01   7.864 1.08e-07 ***
## x6           1.032e-02  1.343e-02   0.769 0.450665
## x7           3.655e-03  1.070e-02   0.342 0.736006
## x8          -1.913e+01  3.197e+01  -0.598 0.555983
## x9           5.052e+01  1.502e+02   0.336 0.739986
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 389.4 on 21 degrees of freedom
## Multiple R-squared:  0.9923, Adjusted R-squared:  0.9889
## F-statistic: 298.9 on 9 and 21 DF,  p-value: < 2.2e-16
```

3. 分别使用前进法，后退法，逐步回归法筛选变量， $= 0.05$

```
lmo31 <- lm(y~1,data=df)
f <- step(lmo31,scope=list(upper=~x1+x2+x3+x4+x5+x6+x7+x8+x9,lower=~1),direction="forward")
```

```
## Start:  AIC=510.41
## y ~ 1
##
##        Df Sum of Sq       RSS    AIC
## + x5    1 364322891  46745970 445.01
## + x1    1 334652105  76416756 460.25
## + x7    1 313570855  97498006 467.80
## + x6    1 253231368 157837493 482.74
## + x3    1 250823965 160244896 483.21
## + x2    1 107641843 303427018 503.00
## + x4    1 100411341 310657520 503.73
## + x9    1  53712849 357356012 508.07
## <none>              411068861 510.41
## + x8    1   6975916 404092945 511.88
##
```

```
## Step:  AIC=445.01
## y ~ x5
##
##        Df Sum of Sq      RSS    AIC
## + x1    1  28070419 18675550 418.57
## + x7    1  13937950 32808020 436.04
## + x3    1   6923191 39822779 442.04
## + x6    1   6210511 40535458 442.59
## <none>              46745970 445.01
## + x2    1   2498398 44247572 445.31
## + x4    1    650568 46095402 446.58
## + x8    1    427015 46318955 446.73
## + x9    1     39461 46706509 446.99
##
## Step:  AIC=418.57
## y ~ x5 + x1
##
##        Df Sum of Sq      RSS    AIC
## + x2    1   9512927  9162624 398.50
## + x7    1   8644999 10030552 401.30
## + x4    1   6911826 11763725 406.24
## + x3    1   4980995 13694556 410.95
## + x6    1   1270214 17405337 418.39
## <none>              18675550 418.57
## + x9    1    308886 18366665 420.05
## + x8    1      2408 18673143 420.57
##
## Step:  AIC=398.5
## y ~ x5 + x1 + x2
##
##        Df Sum of Sq     RSS    AIC
## + x3    1   5717883 3444741 370.17
## + x7    1   2466099 6696525 390.78
## + x4    1   1576284 7586340 394.64
## <none>              9162624 398.50
## + x8    1    310651 8851972 399.43
## + x6    1     84360 9078264 400.21
## + x9    1      5375 9157248 400.48
##
```

```
## Step:  AIC=370.17
## y ~ x5 + x1 + x2 + x3
## 
##        Df Sum of Sq     RSS    AIC
## <none>              3444741 370.17
## + x6    1    137540 3307201 370.91
## + x7    1     89068 3355673 371.36
## + x8    1     55576 3389165 371.67
## + x9    1      4674 3440066 372.13
## + x4    1         3 3444737 372.17
```

```
summary(f)
```

```
## 
## Call:
## lm(formula = y ~ x5 + x1 + x2 + x3, data = df)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -943.18 -161.05   12.74  250.93  566.25 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -1694.6269   562.9773  -3.010  0.00574 ** 
## x5              1.7424     0.1912   9.111 1.42e-09 ***
## x1              1.3642     0.0861  15.844 7.11e-15 ***
## x2              1.7679     0.2010   8.796 2.86e-09 ***
## x3              2.2894     0.3485   6.569 5.76e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 364 on 26 degrees of freedom
## Multiple R-squared:  0.9916, Adjusted R-squared:  0.9903 
## F-statistic: 769.2 on 4 and 26 DF,  p-value: < 2.2e-16
```

```
b <- step(olsa,direction='backward',trace = 1)
```

```
## Start:  AIC=377.73
## y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9
## 
```

```
##        Df Sum of Sq      RSS    AIC
## - x4    1        21  3184326 375.73
## - x9    1     17149  3201454 375.90
## - x7    1     17700  3202005 375.90
## - x8    1     54295  3238599 376.26
## - x6    1     89586  3273891 376.59
## <none>               3184305 377.73
## - x3    1   2662593  5846898 394.57
## - x2    1   4561056  7745361 403.29
## - x5    1   9377500 12561805 418.28
## - x1    1  23314547 26498852 441.42
##
## Step:  AIC=375.73
## y ~ x1 + x2 + x3 + x5 + x6 + x7 + x8 + x9
##
##        Df Sum of Sq      RSS    AIC
## - x9    1     17428  3201754 373.90
## - x7    1     18563  3202889 373.91
## - x8    1     54437  3238763 374.26
## - x6    1     91813  3276139 374.61
## <none>               3184326 375.73
## - x3    1   2936130  6120456 393.99
## - x2    1   5467941  8652267 404.72
## - x5    1   9393345 12577671 416.32
## - x1    1  25886086 29070412 442.29
##
## Step:  AIC=373.9
## y ~ x1 + x2 + x3 + x5 + x6 + x7 + x8
##
##        Df Sum of Sq      RSS    AIC
## - x7    1     34634  3236387 372.24
## - x6    1     74800  3276554 372.62
## - x8    1     82150  3283904 372.69
## <none>               3201754 373.90
## - x3    1   3055353  6257107 392.67
## - x2    1   5725836  8927590 403.69
## - x5    1   9382624 12584378 414.33
## - x1    1  25868832 29070586 440.29
##
```

```
## Step:  AIC=372.24
## y ~ x1 + x2 + x3 + x5 + x6 + x8
##
##        Df Sum of Sq      RSS    AIC
## - x8    1     70813  3307201 370.91
## - x6    1    152777  3389165 371.67
## <none>               3236387 372.24
## - x3    1   5501284  8737672 401.02
## - x2    1   8895049 12131436 411.20
## - x5    1   9458098 12694485 412.60
## - x1    1  27733098 30969486 440.25
##
## Step:  AIC=370.91
## y ~ x1 + x2 + x3 + x5 + x6
##
##        Df Sum of Sq      RSS    AIC
## - x6    1    137540  3444741 370.17
## <none>               3307201 370.91
## - x3    1   5771063  9078264 400.21
## - x2    1   8871193 12178394 409.32
## - x5    1   9473521 12780722 410.81
## - x1    1  28248162 31555363 438.83
##
## Step:  AIC=370.17
## y ~ x1 + x2 + x3 + x5
##
##        Df Sum of Sq      RSS    AIC
## <none>               3444741 370.17
## - x3    1   5717883  9162624 398.50
## - x2    1  10249815 13694556 410.95
## - x5    1  10998313 14443054 412.60
## - x1    1  33258637 36703378 441.52
```

```
summary(b)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x5, data = df)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max 
## -943.18 -161.05   12.74  250.93  566.25 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -1694.6269   562.9773  -3.010  0.00574 ** 
## x1              1.3642     0.0861  15.844 7.11e-15 ***
## x2              1.7679     0.2010   8.796 2.86e-09 ***
## x3              2.2894     0.3485   6.569 5.76e-07 ***
## x5              1.7424     0.1912   9.111 1.42e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 364 on 26 degrees of freedom
## Multiple R-squared:  0.9916, Adjusted R-squared:  0.9903 
## F-statistic: 769.2 on 4 and 26 DF,  p-value: < 2.2e-16
```

```
s <- step(olsa,direction='both',trace = 1)
```

```
## Start:  AIC=377.73
## y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9
## 
##        Df Sum of Sq      RSS    AIC
## - x4    1        21  3184326 375.73
## - x9    1     17149  3201454 375.90
## - x7    1     17700  3202005 375.90
## - x8    1     54295  3238599 376.26
## - x6    1     89586  3273891 376.59
## <none>               3184305 377.73
## - x3    1   2662593  5846898 394.57
## - x2    1   4561056  7745361 403.29
## - x5    1   9377500 12561805 418.28
## - x1    1  23314547 26498852 441.42
## 
## Step:  AIC=375.73
## y ~ x1 + x2 + x3 + x5 + x6 + x7 + x8 + x9
## 
##        Df Sum of Sq      RSS    AIC
## - x9    1     17428  3201754 373.90
## - x7    1     18563  3202889 373.91
## - x8    1     54437  3238763 374.26
## - x6    1     91813  3276139 374.61
## <none>                3184326 375.73
## + x4    1        21  3184305 377.73
## - x3    1   2936130  6120456 393.99
## - x2    1   5467941  8652267 404.72
## - x5    1   9393345 12577671 416.32
## - x1    1  25886086 29070412 442.29
##
## Step:  AIC=373.9
## y ~ x1 + x2 + x3 + x5 + x6 + x7 + x8
##
##         Df Sum of Sq      RSS    AIC
## - x7    1     34634  3236387 372.24
## - x6    1     74800  3276554 372.62
## - x8    1     82150  3283904 372.69
## <none>                3201754 373.90
## + x9    1     17428  3184326 375.73
## + x4    1       300  3201454 375.90
## - x3    1   3055353  6257107 392.67
## - x2    1   5725836  8927590 403.69
## - x5    1   9382624 12584378 414.33
## - x1    1  25868832 29070586 440.29
##
## Step:  AIC=372.24
## y ~ x1 + x2 + x3 + x5 + x6 + x8
##
##         Df Sum of Sq      RSS    AIC
## - x8    1     70813  3307201 370.91
## - x6    1    152777  3389165 371.67
## <none>                3236387 372.24
## + x7    1     34634  3201754 373.90
## + x9    1     33499  3202889 373.91
## + x4    1       704  3235683 374.23
## - x3    1   5501284  8737672 401.02
## - x2    1   8895049 12131436 411.20
## - x5    1   9458098 12694485 412.60
## - x1    1  27733098 30969486 440.25
##
```

```
## Step:  AIC=370.91
## y ~ x1 + x2 + x3 + x5 + x6
##
##        Df Sum of Sq      RSS    AIC
## - x6    1    137540  3444741 370.17
## <none>               3307201 370.91
## + x8    1     70813  3236387 372.24
## + x9    1     60187  3247014 372.34
## + x7    1     23297  3283904 372.69
## + x4    1         2  3307199 372.91
## - x3    1   5771063  9078264 400.21
## - x2    1   8871193 12178394 409.32
## - x5    1   9473521 12780722 410.81
## - x1    1  28248162 31555363 438.83
##
## Step:  AIC=370.17
## y ~ x1 + x2 + x3 + x5
##
##        Df Sum of Sq      RSS    AIC
## <none>               3444741 370.17
## + x6    1    137540  3307201 370.91
## + x7    1     89068  3355673 371.36
## + x8    1     55576  3389165 371.67
## + x9    1      4674  3440066 372.13
## + x4    1         3  3444737 372.17
## - x3    1   5717883  9162624 398.50
## - x2    1  10249815 13694556 410.95
## - x5    1  10998313 14443054 412.60
## - x1    1  33258637 36703378 441.52
```

```
summary(s)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x5, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -943.18 -161.05   12.74  250.93  566.25
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1694.6269   562.9773  -3.010  0.00574 **
## x1              1.3642     0.0861  15.844 7.11e-15 ***
## x2              1.7679     0.2010   8.796 2.86e-09 ***
## x3              2.2894     0.3485   6.569 5.76e-07 ***
## x5              1.7424     0.1912   9.111 1.42e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 364 on 26 degrees of freedom
## Multiple R-squared:  0.9916, Adjusted R-squared:  0.9903
## F-statistic: 769.2 on 4 and 26 DF,  p-value: < 2.2e-16
```