Homework 8

陈旭鹏

2018/4/20

1

a

```
# read the data
setwd('~/Desktop/三春/5线性回归分析/作业/HW8/')
dat<-read.csv("hw8.csv")
X1<-dat$x1
X2<-dat$x2
X3<-dat$x3
X4<-dat$x4
Y<-dat$y
# plot stem and leaf plots
stem(X1)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##    6 | 248
##    8 | 4671468
##   10 | 014456902
##   12 | 0003
##   14 | 00
```

```
stem(X2)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##    6 | 37
##    8 | 135947
##   10 | 127034789
##   12 | 01112599
```

```
stem(X3)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##    8 | 0
##    9 | 01335556789
##   10 | 002356789
##   11 | 3456
```

```
stem(X4)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##    7 | 48
##    8 | 03457889
##    9 | 0557
##   10 | 0223345889
##   11 | 0
```

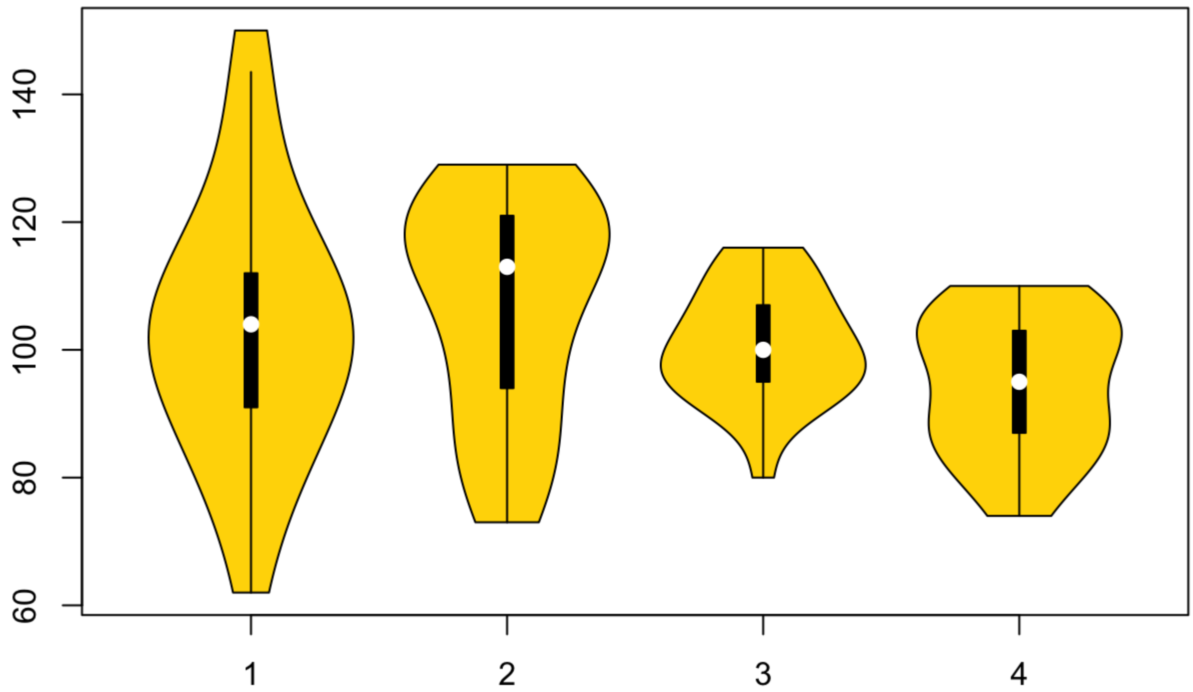
It seems that X3 has a denser concentration and the following boxplot supports it. X1 has two outliers. X2 is asymmetric

```
library(vioplot)
```

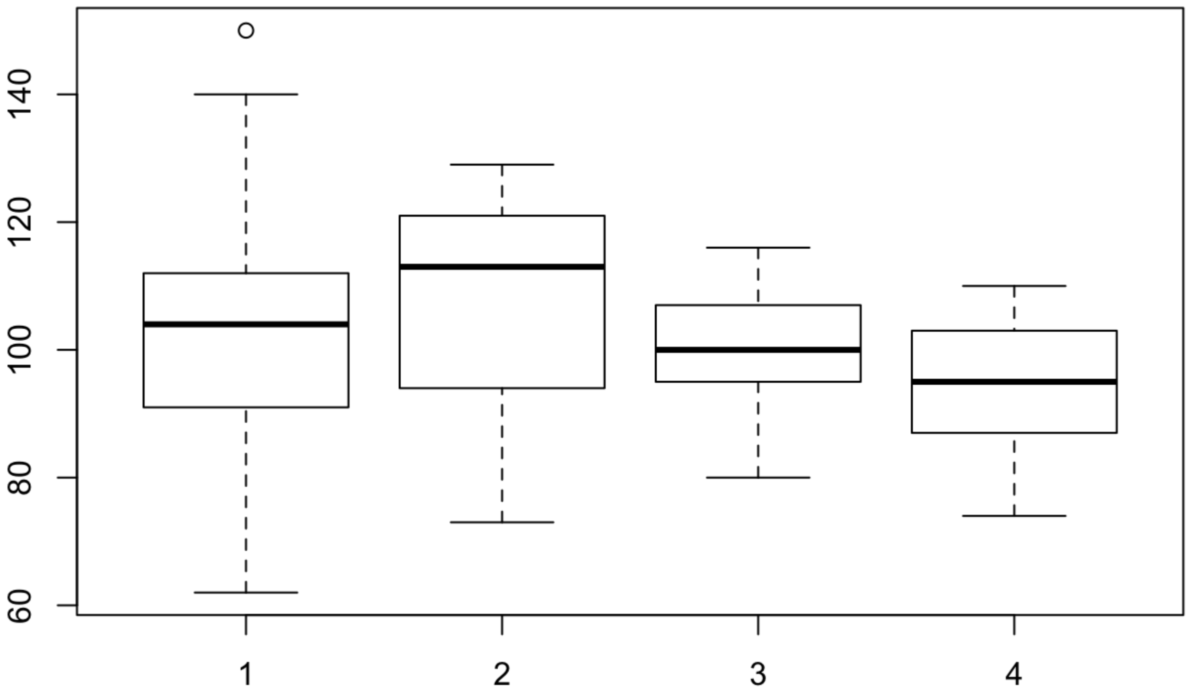
```
## Loading required package: sm
```

```
## Package 'sm', version 2.2-5.4: type help(sm) for summary information
```

```
vioplot(X1,X2,X3,X4,col="gold")
```

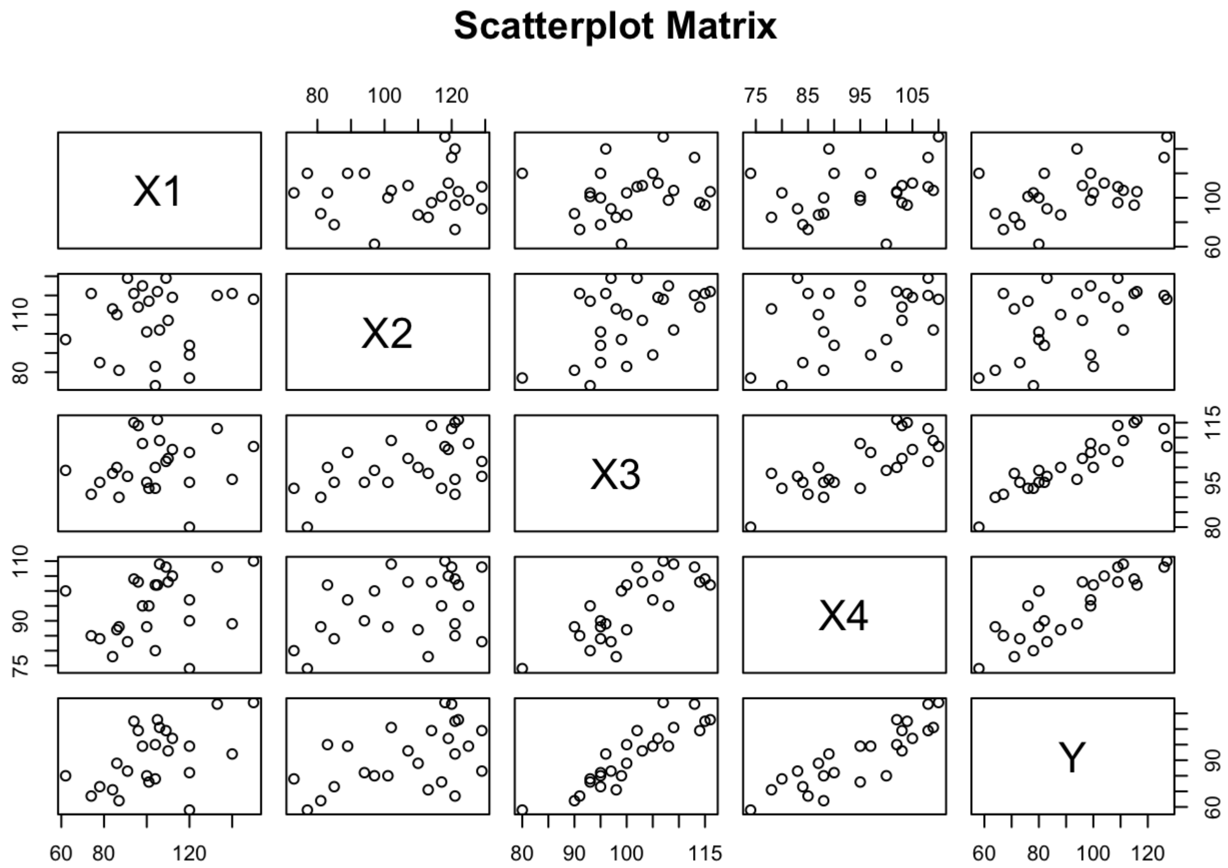



```
boxplot(X1,X2,X3,X4)
```



b

```
pairs(~X1+X2+X3+X4+Y,data=dat,
   main="Scatterplot Matrix")
```




```
cor(dat)
```

```
##            y        x1        x2        x3        x4
## y  1.0000000 0.5144107 0.4970057 0.8970645 0.8693865
## x1 0.5144107 1.0000000 0.1022689 0.1807692 0.3266632
## x2 0.4970057 0.1022689 1.0000000 0.5190448 0.3967101
## x3 0.8970645 0.1807692 0.5190448 1.0000000 0.7820385
## x4 0.8693865 0.3266632 0.3967101 0.7820385 1.0000000
```

obviously X3 and X4 has high correlation.

C

```
Fit = lm(Y~X1+X2+X3+X4, data=dat)
anova(Fit)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X1         1 2395.9  2395.9 142.620 1.480e-10 ***
## X2         1 1807.0  1807.0 107.565 1.708e-09 ***
## X3         1 4254.5  4254.5 253.259 8.045e-13 ***
## X4         1  260.7   260.7  15.521   0.00081 ***
## Residuals 20  336.0    16.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(Fit)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9779 -3.4506  0.0941  2.4749  5.9959
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -124.38182    9.94106 -12.512 6.48e-11 ***
## X1             0.29573    0.04397   6.725 1.52e-06 ***
## X2             0.04829    0.05662   0.853  0.40383
## X3             1.30601    0.16409   7.959 1.26e-07 ***
## X4             0.51982    0.13194   3.940  0.00081 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.099 on 20 degrees of freedom
## Multiple R-squared:  0.9629, Adjusted R-squared:  0.9555
## F-statistic: 129.7 on 4 and 20 DF,  p-value: 5.262e-14
```

$\hat{Y} = -124.38 + 0.30x_1 + 0.05x_2 + 1.31x_3 + 0.52x_4$ It seems X2 should be excluded from the model since the p-value=0.4038.

2

a

```
library(leaps)
best <- function(model, ...)
{
  subsets <- regsubsets(formula(model), model.frame(model), ...)
  subsets <- with(summary(subsets),
                  cbind(p = as.numeric(rownames(which)), which, adjr2))

  return(subsets)
}
round(best(Fit, nbest = 6), 4)
```

```
##   p (Intercept) X1 X2 X3 X4  adjr2
## 1 1           1  0  0  1  0 0.7962
## 1 1           1  0  0  0  1 0.7452
## 1 1           1  1  0  0  0 0.2326
## 1 1           1  0  1  0  0 0.2143
## 2 2           1  1  0  1  0 0.9269
## 2 2           1  0  0  1  1 0.8661
## 2 2           1  1  0  0  1 0.7985
## 2 2           1  0  1  1  0 0.7884
## 2 2           1  0  1  0  1 0.7636
## 2 2           1  1  1  0  0 0.4155
## 3 3           1  1  0  1  1 0.9560
## 3 3           1  1  1  1  0 0.9247
## 3 3           1  0  1  1  1 0.8617
## 3 3           1  1  1  0  1 0.8233
## 4 4           1  1  1  1  1 0.9555
```

The four best subset regression models are

subset	$R^2_{a,p}$
x1, x3, x4	0.956
x1,x2,x3,x4	0.955
x1,x3	0.927
x1,x2,x3	0.925

b

There are C_p Criterion, #AIC_p# and #SBC_p# which can be used as criterion to select the best model. They all place penalties for adding predictors.

3

```
library(MASS)
```

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

```
## The following object is masked from 'package:sm':
##
##     muscle
```

```
Null = lm(Y ~ 1, dat)
```

```
addterm(Null, scope = Fit, test="F")
```

```
## Single term additions
##
## Model:
## Y ~ 1
##        Df Sum of Sq    RSS    AIC F Value     Pr(F)
## <none>              9054.0 149.30
## X1      1    2395.9 6658.1 143.62   8.276  0.008517 **
## X2      1    2236.5 6817.5 144.21   7.545  0.011487 *
## X3      1    7286.0 1768.0 110.47  94.782 1.264e-09 ***
## X4      1    6843.3 2210.7 116.06  71.198 1.699e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
NewMod = update( Null, .~. + X3)
addterm( NewMod, scope = Fit, test="F" )
```

```
## Single term additions
##
## Model:
## Y ~ X3
##        Df Sum of Sq     RSS     AIC F Value     Pr(F)
## <none>              1768.02 110.469
## X1      1   1161.37  606.66  85.727  42.116 1.578e-06 ***
## X2      1     12.21 1755.81 112.295   0.153   0.69946
## X4      1    656.71 1111.31 100.861  13.001   0.00157 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
NewMod = update( NewMod, .~. + X1)
dropterm(NewMod , test = "F")
```

```
## Single term deletions
##
## Model:
## Y ~ X3 + X1
##        Df Sum of Sq    RSS     AIC F Value     Pr(F)
## <none>               606.7  85.727
## X3      1    6051.5 6658.1 143.618 219.453 6.313e-13 ***
## X1      1    1161.4 1768.0 110.469  42.116 1.578e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
addterm( NewMod, scope = Fit, test="F" )
```

```
## Single term additions
##
## Model:
## Y ~ X3 + X1
##        Df Sum of Sq    RSS    AIC F Value     Pr(F)
## <none>               606.66 85.727
## X2      1     9.937 596.72 87.314  0.3497 0.5605965
## X4      1   258.460 348.20 73.847 15.5879 0.0007354 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
NewMod = update( NewMod, .~. + X4)
dropterm( NewMod, test = "F" )
```

```
## Single term deletions
##
## Model:
## Y ~ X3 + X1 + X4
##        Df Sum of Sq     RSS     AIC F Value     Pr(F)
## <none>                348.20  73.847
## X3      1   1324.39 1672.59 111.081  79.875 1.334e-08 ***
## X1      1    763.12 1111.31 100.861  46.024 1.040e-06 ***
## X4      1    258.46  606.66  85.727  15.588 0.0007354 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
addterm( NewMod, scope = Fit, test="F" )
```

```
## Single term additions
##
## Model:
## Y ~ X3 + X1 + X4
##        Df Sum of Sq    RSS    AIC F Value  Pr(F)
## <none>               348.20 73.847
## X2      1     12.22 335.98 74.954  0.7274 0.4038
```

- As shown, start with no predictors, X3 is chosen because of smallest p-value.
- Then regressing y on x3 and additional one predictor, the result shows that X1 has the smallest p-value (1.578e-06< 0.05). Therefore X1 can be included in the model. In the same time a test is given to see if x3 should be dtropped. Since p-value (6.313e-13<0.10), X3 is retained.
- Then regressing y on X3, X1 and any one of the rest two, it shows that X4 has the smallest p-value (0.0007354 < 0.05) and hence being included in the model.In the same time a test is given to see if x3 or x1 should be dtropped. Since both of their p-value < 0.10, they are both retained.
- Finally, regressing y on all four predictors and x2 isn’t significant to be included (0.4038 > 0.05). Thus it is deleted from the model.
- The best subset of predictor variables to predict job proficiency is (x1,x3,x4)

b

The model evaluated using the forward stepwise regression shows the same result as earlier chosen variables under the criteria of adjusted R square.