Homework 9

陈旭鹏

2018/5/2

1

```
# read the data
setwd('~/Desktop/三春/5线性回归分析/作业/HW9/')
dat<-read.csv("hw6.csv")
cases<-dat$X1
percent<-dat$X2
holiday<-dat$X3
labor<-dat$Y
```

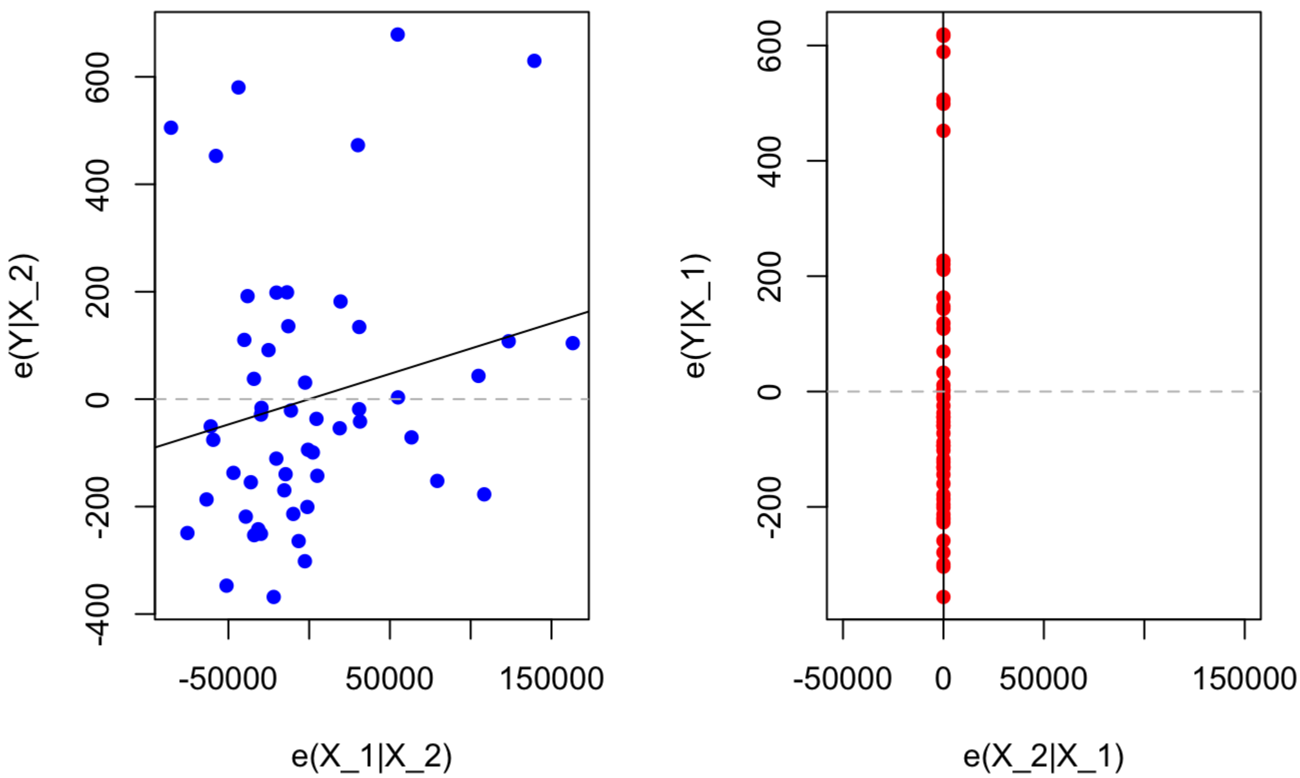
a

```
dat.fit<-lm(labor~cases+percent)
sm<-summary(dat.fit)
sm
```

```
##
## Call:
## lm(formula = labor ~ cases + percent)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -347.54 -160.95  -52.52  107.56  627.11
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.106e+03  3.054e+02  13.444   <2e-16 ***
## cases        9.425e-04  6.327e-04   1.490    0.143
## percent     -4.054e+00  3.710e+01  -0.109    0.913
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 248.5 on 49 degrees of freedom
## Multiple R-squared:  0.04336,    Adjusted R-squared:  0.004311
## F-statistic:  1.11 on 2 and 49 DF,  p-value: 0.3376
```

b

```
yonx1 <- lm(labor~cases,data=dat)
yonx2 <- lm(labor~percent,data=dat)
x1onx2 <- lm(cases~percent,data=dat)
x2onx1 <- lm(percent~cases,data=dat)
par(mfrow=c(1,2))
plot(x1onx2$residuals,yonx2$residuals,col="blue",pch=16,
xlab="e(X_1|X_2)", ylab="e(Y|X_2)")
abline(0,dat.fit$coefficients[2])
abline(0,0,lty=2,col="gray")
plot(x2onx1$residuals,yonx1$residuals,col="red",pch=16,
xlab="e(X_2|X_1)", ylab="e(Y|X_1)",xlim=c(-50000,150000))
abline(0,dat.fit$coefficients[3])
abline(0,0,lty=2,col="gray")
```



From the boxplot, we can know the median, maximum, minimum, 25 and 75 percent quantile of the residuals.

c

We use about the same scale in the two plots. In the first plot, the scatter of points around the least square line does not differ much compared to the scatter around the horizontal line. However, in the second plot, we can see that the scatter around the regression line (which is almost verticle under this scale) is significantly smaller than the scatter around the horizontal line. This tells us that X1 is of little use when X2 is in the model, while X2 can still explain a lot when X1 is present. So perhaps X1 can be discarded.

d

```r
# The regression functions below are required
fit1<-lm(resid(yonx2)~resid(x1onx2)-1)
summary(fit1)
```

```
##
## Call:
## lm(formula = resid(yonx2) ~ resid(x1onx2) - 1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -347.54 -160.95  -52.52  107.56  627.11
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## resid(x1onx2) 0.0009425  0.0006201    1.52    0.135
##
## Residual standard error: 243.5 on 51 degrees of freedom
## Multiple R-squared:  0.04333,    Adjusted R-squared:  0.02457
## F-statistic:  2.31 on 1 and 51 DF,  p-value: 0.1347
```

```r
summary(yonx2)
```

```
##
## Call:
## lm(formula = labor ~ percent, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -368.22 -171.59  -46.21  108.41  678.78
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4352.671    259.566   16.77   <2e-16 ***
## percent        1.506     37.360    0.04    0.968
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 251.5 on 50 degrees of freedom
## Multiple R-squared:  3.249e-05,  Adjusted R-squared:  -0.01997
## F-statistic: 0.001625 on 1 and 50 DF,  p-value: 0.968
```

```r
summary(x1onx2)
```

```
##
## Call:
## lm(formula = cases ~ percent, data = dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -85531 -34666 -13313  22128 163203
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   262080      57328   4.572  3.2e-05 ***
## percent         5899       8252   0.715    0.478
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 55540 on 50 degrees of freedom
## Multiple R-squared:  0.01012,    Adjusted R-squared:  -0.009679
## F-statistic: 0.5111 on 1 and 50 DF,  p-value: 0.478
```

2

a

```
lm.b<-lm(labor~cases+percent+holiday)
(rsd.lm=round(rstudent(lm.b), 3))
```

```
##      1      2      3      4      5      6      7      8      9     10
## -0.197  1.243 -0.174 -0.345  0.591  0.162 -1.090 -1.174 -0.971  2.043
##     11     12     13     14     15     16     17     18     19     20
##  0.975 -0.221 -0.458 -1.581  0.142 -0.960  1.252 -0.536  0.741 -1.326
##     21     22     23     24     25     26     27     28     29     30
## -0.384  0.120 -0.837  0.575  0.385 -0.146  0.562 -0.390  0.412 -0.192
##     31     32     33     34     35     36     37     38     39     40
## -1.024 -1.992  1.643  1.725 -1.698 -0.446 -1.016  2.061 -0.755  2.170
##     41     42     43     44     45     46     47     48     49     50
##  0.164  0.613  0.864  0.451  0.314 -0.590 -1.027 -0.236  0.025  1.568
##     51     52
## -1.414  0.423
```

```
n=25
p=4
ifelse(rsd.lm > qt(1-0.95/2/n,n-p-1), "outlier", "Non-outlier")
```

```
##             1             2             3             4             5
## "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier"
##             6             7             8             9            10
## "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier"
##            11            12            13            14            15
## "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier"
##            16            17            18            19            20
## "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier"
##            21            22            23            24            25
## "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier"
##            26            27            28            29            30
## "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier"
##            31            32            33            34            35
## "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier"
##            36            37            38            39            40
## "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier"
##            41            42            43            44            45
## "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier"
##            46            47            48            49            50
## "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier"
##            51            52
## "Non-outlier" "Non-outlier"
```

It appears to be that all observed values cannot be definited as outliers by Bonferroni outlier test.

b

```
(h.lm=round(hatvalues(lm.b), 3))
```

```
##     1     2     3     4     5     6     7     8     9    10    11    12
## 0.022 0.045 0.210 0.049 0.205 0.039 0.029 0.044 0.038 0.041 0.029 0.041
##    13    14    15    16    17    18    19    20    21    22    23    24
## 0.040 0.050 0.045 0.259 0.042 0.051 0.022 0.022 0.193 0.244 0.058 0.073
##    25    26    27    28    29    30    31    32    33    34    35    36
## 0.072 0.022 0.027 0.075 0.044 0.036 0.028 0.125 0.060 0.025 0.039 0.047
##    37    38    39    40    41    42    43    44    45    46    47    48
## 0.039 0.023 0.060 0.031 0.059 0.123 0.271 0.225 0.123 0.039 0.064 0.282
##    49    50    51    52
## 0.025 0.023 0.100 0.023
```

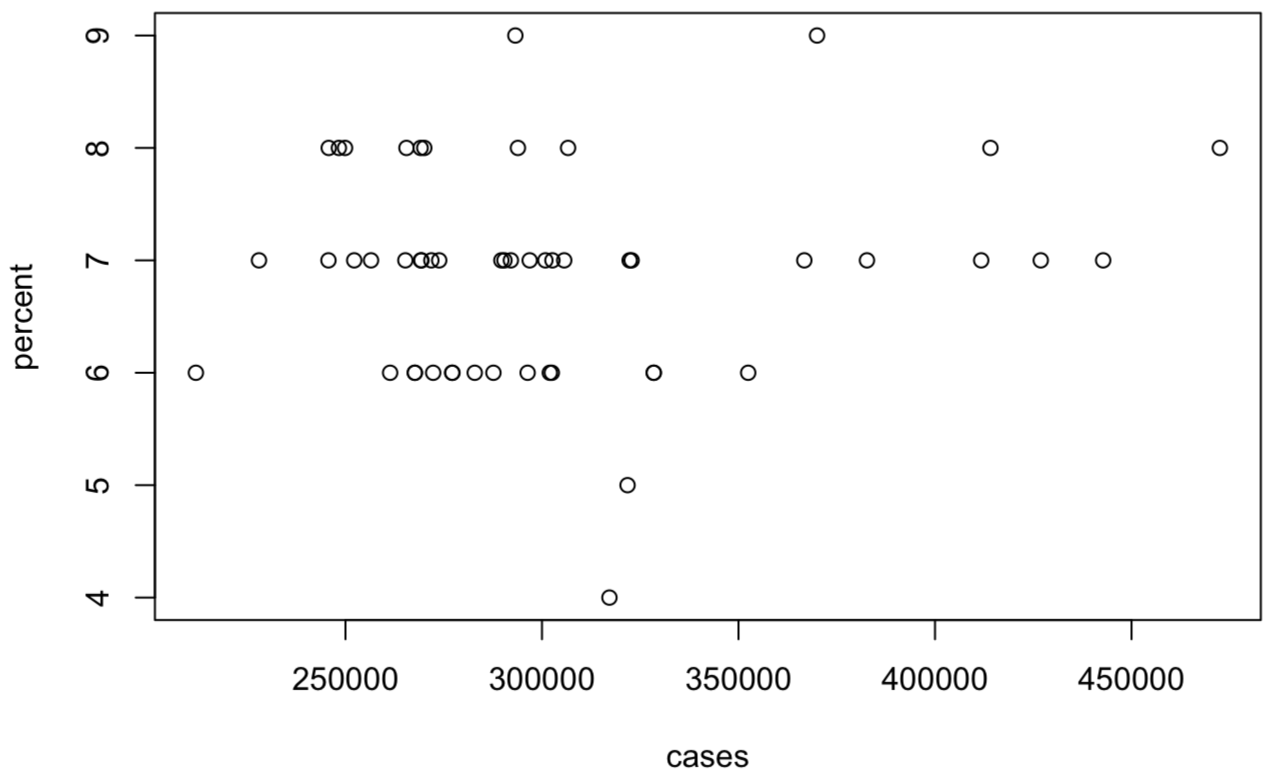
```
n=25
p=4
ifelse(h.lm > 2*p/n, "outlier", "Non-outlier")
```

```
##             1             2             3             4             5 
## "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" 
##             6             7             8             9            10 
## "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" 
##            11            12            13            14            15 
## "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" 
##            16            17            18            19            20 
## "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" 
##            21            22            23            24            25 
## "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" 
##            26            27            28            29            30 
## "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" 
##            31            32            33            34            35 
## "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" 
##            36            37            38            39            40 
## "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" 
##            41            42            43            44            45 
## "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" 
##            46            47            48            49            50 
## "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" 
##            51            52 
## "Non-outlier" "Non-outlier"
```

It appears to be that all observed values cannot be definited as outliers by the rule of thumb.

C

```
attach(dat)
plot(cases,percent)
```



Judging from the scatter plot, this prediction does not seem to involve extrapolation beyond the range of the data.

```
xnew<-c(300000,7.2,0)
X<-as.matrix(dat)
X<-X[,-1]
hnew<-t(xnew)%*%solve(t(X)%*%X)%*%xnew
ifelse(hnew > 2*p/n, "YES", "NO")
```

```
##      [,1]
## [1,] "NO"
```

d

```
a= cbind(
  "DFFITS"  = round(dffits(lm.b), 4),
  "DFBETA0" = round(dfbetas( lm.b)[,1], 4),
  "DFBETA1" = round(dfbetas( lm.b)[,2], 4),
  "DFBETA2" = round(dfbetas( lm.b)[,3], 4),
  "DFBETA3" = round(dfbetas( lm.b)[,4], 4),
  "Cook's D" = round(cooks.distance( lm.b), 4))
a[c(16,22,43,48),]
```

```
##     DFFITS DFBETA0 DFBETA1 DFBETA2 DFBETA3 Cook's D
## 16 -0.5670 -0.2388 -0.0674  0.3365 -0.4391   0.0805
## 22  0.0684  0.0347 -0.0324 -0.0173  0.0554   0.0012
## 43  0.5264 -0.3184  0.1258  0.2871  0.3692   0.0696
## 48 -0.1478  0.0513 -0.0945  0.0093 -0.1028   0.0056
```

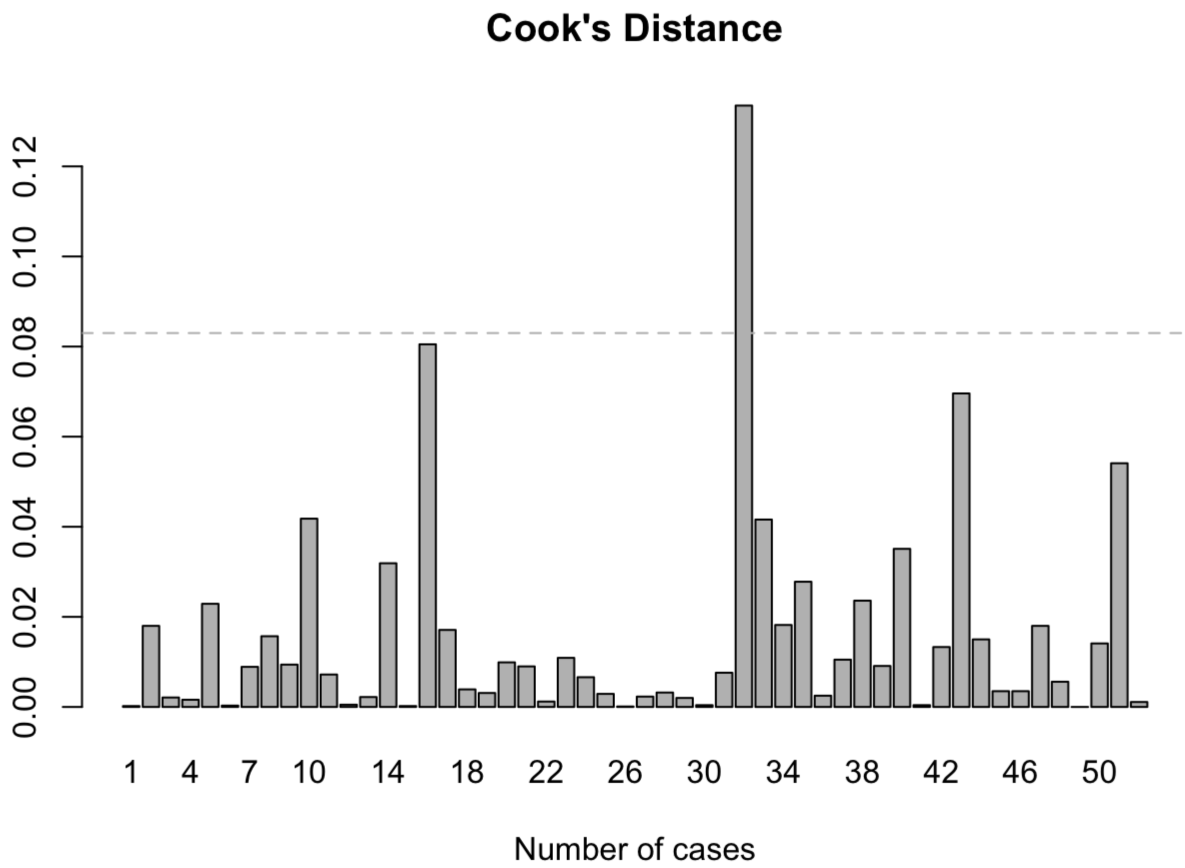
```
a[c(10,32,38, 40),]
```

```
##     DFFITS DFBETA0 DFBETA1 DFBETA2 DFBETA3 Cook's D
## 10  0.4221  0.3175 -0.1038 -0.2582 -0.0880   0.0418
## 32 -0.7529  0.4615  0.1135 -0.6827  0.1319   0.1335
## 38  0.3176  0.0404 -0.0684  0.0470 -0.1032   0.0236
## 40  0.3890  0.1173 -0.2096  0.0632 -0.1034   0.0351
```

- observation is deemed influential if the absolute value of its Cook's Distance value is greater than: $4/(N-k-1) = 0.083$ so case32 should be considered influential.
- observation is deemed influential if the absolute value of its DFBETAS value is greater than: $\frac{2}{\sqrt{n}} = 0.27$, so it seem that 16 and 32 are influential, but 22 and 48 seem non-influential.
- since An observation is deemed influential if the absolute value of its DFFITS value is greater than: $2 * \frac{\sqrt{(p+1)}}{(n-p-1)} = 0.083$, so 10,32,38,40 all sesm influential.

f

```
a= ("Cook's D" = round(cooks.distance( lm.b), 4))
barplot(a[seq(1, 52)], main="Cook's Distance",
    xlab="Number of cases")
abline(0.083,0,lty=2,col="gray")
```

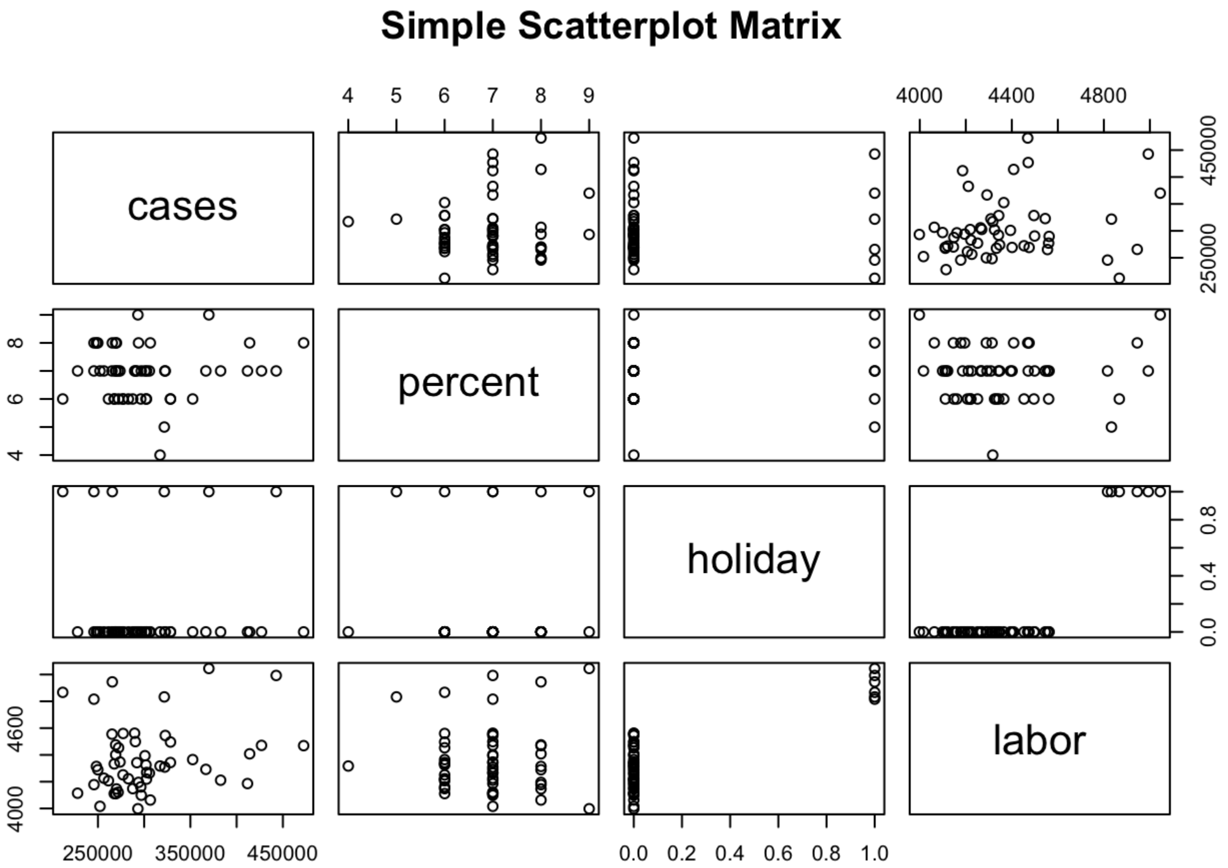


So case 32 is influential.

3

a

```
pairs(~cases+percent+holiday+labor,data=dat,
   main="Simple Scatterplot Matrix")
```




```
cor(dat[,1:4])
```

```
##              Y          X1          X2         X3
## Y  1.000000000 0.20766494 0.005700383 0.81057940
## X1 0.207664935 1.00000000 0.100592161 0.04565698
## X2 0.005700383 0.10059216 1.000000000 0.04464371
## X3 0.810579396 0.04565698 0.044643714 1.00000000
```

It seems that $X_1 and X_3$ have the strongest linear associations.

b

```
#cases+percent+holiday
summary(lm(cases ~ percent+holiday))$r.squared
```

```
## [1] 0.01181682
```

```
summary(lm(percent~cases +holiday))$r.squared
```

```
## [1] 0.01172621
```

```
summary(lm(holiday~cases +percent))$r.squared
```

```
## [1] 0.003705038
```

```r
myfun<-function(a){
  result <-1/(1-a**2)
  return (result)
  }
myfun(0.01181682)
```

```
## [1] 1.00014
```

```r
myfun(0.01172621)
```

```
## [1] 1.000138
```

```r
myfun(0.003705038)
```

```
## [1] 1.000014
```

$$(VIF)_j = \frac{1}{1 - R_j^2}$$

$$max_j (VIF)_j = 1.00014 \leq 10$$

So there is no serious multicolinearity here.