Homework 6

陈旭鹏

2018/4/8

1

```
# read the data
setwd('~/Desktop/三春/5线性回归分析/作业/HW6/')
dat<-read.csv("hw6.csv")
cases<-dat$X1
percent<-dat$X2
holiday<-dat$X3
labor<-dat$Y
# plot stem and leaf plots
stem(cases)
```

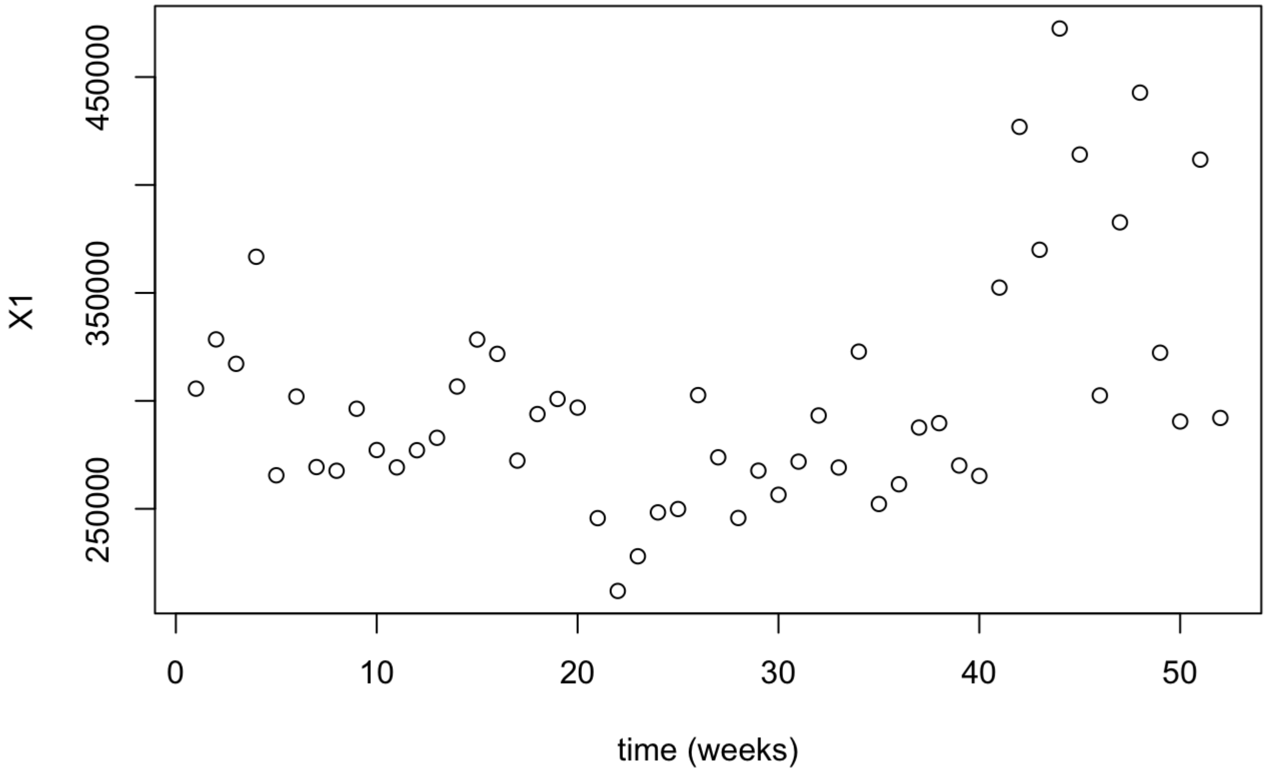
```
##
##   The decimal point is 5 digit(s) to the right of the |
##
##   2 | 13
##   2 | 5555566777777777788899999
##   3 | 00000011222233
##   3 | 5778
##   4 | 1134
##   4 | 7
```

```
stem(percent)
```

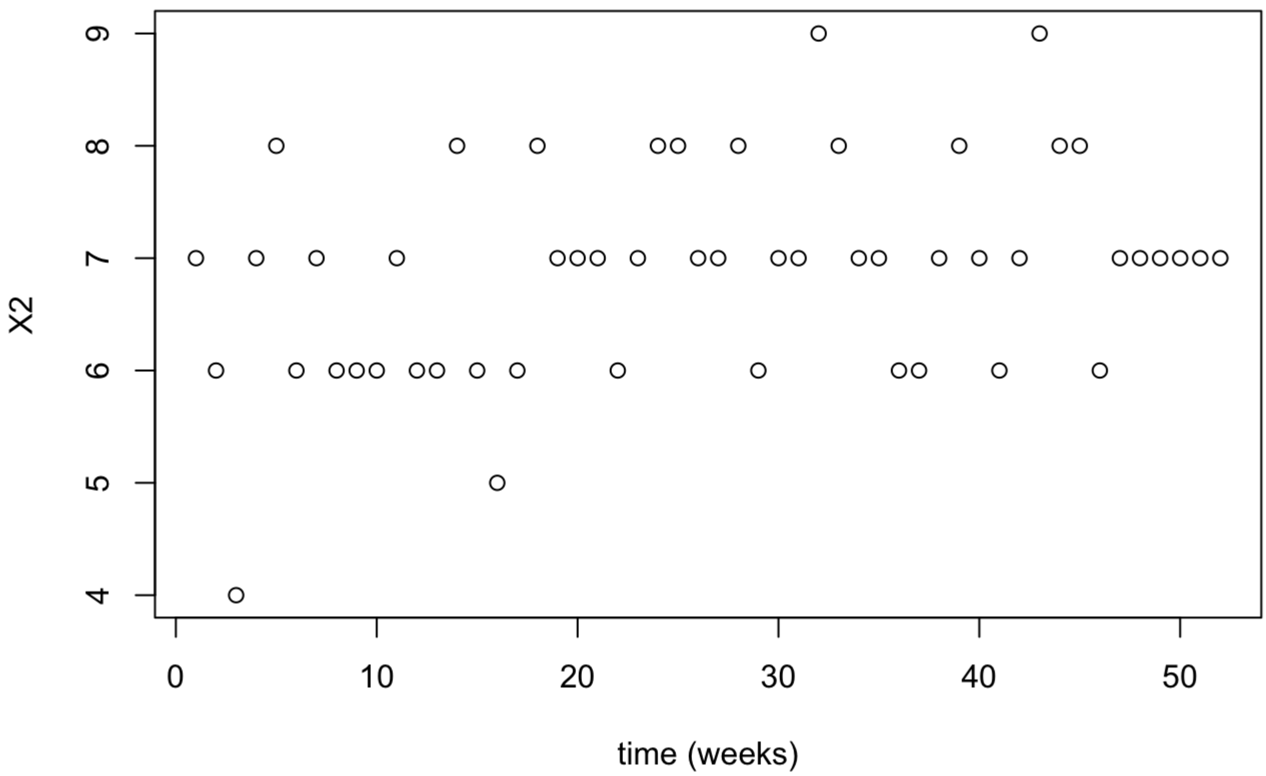
```
##
##   The decimal point is at the |
##
##   4 | 0
##   5 | 0
##   6 | 000000000000000
##   7 | 00000000000000000000000
##   8 | 0000000000
##   9 | 00
```

The plots are above. There seems to be some outliers. For example for X_2, the values more than 9 and lower than 5 seem to be outliers. The gaps are obvious.

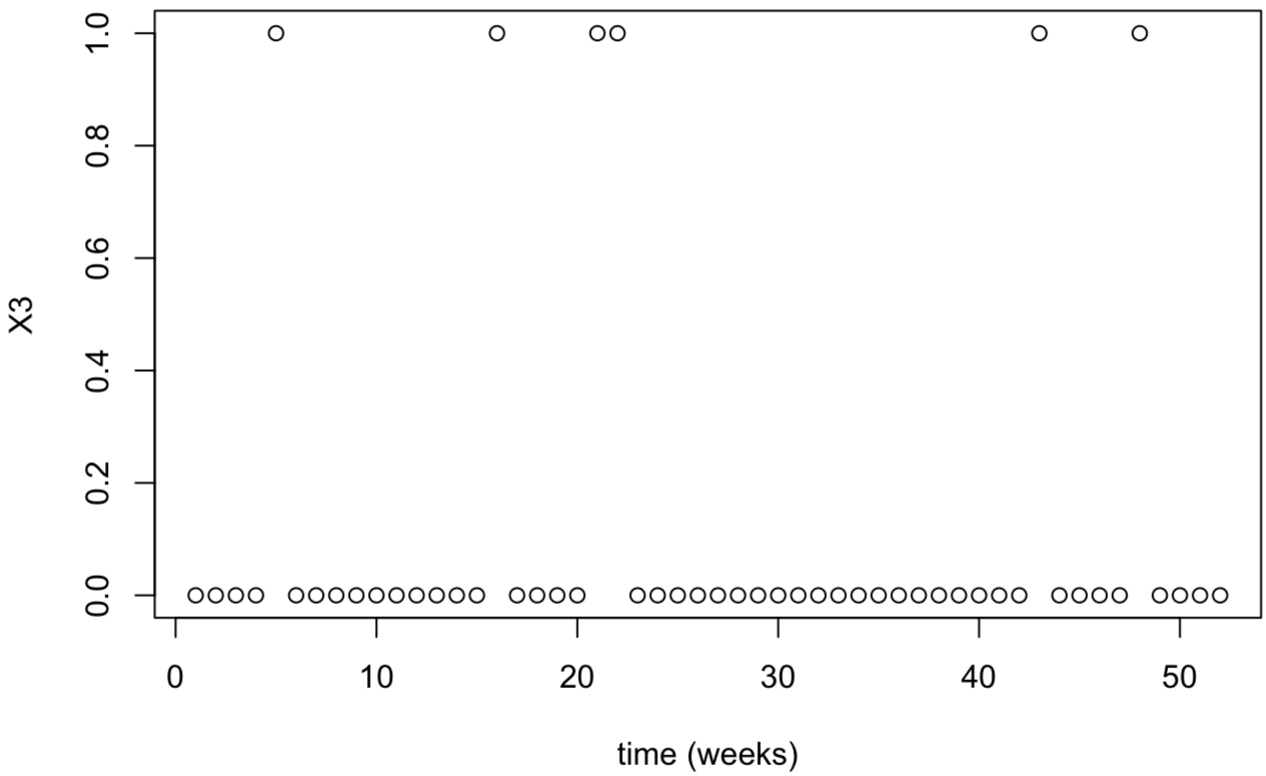
```
# plot time plots
time<-1:52
plot(time,dat$X1,xlab="time (weeks)",ylab="X1")
```




```
plot(time,dat$X2,xlab="time (weeks)",ylab="X2")
```

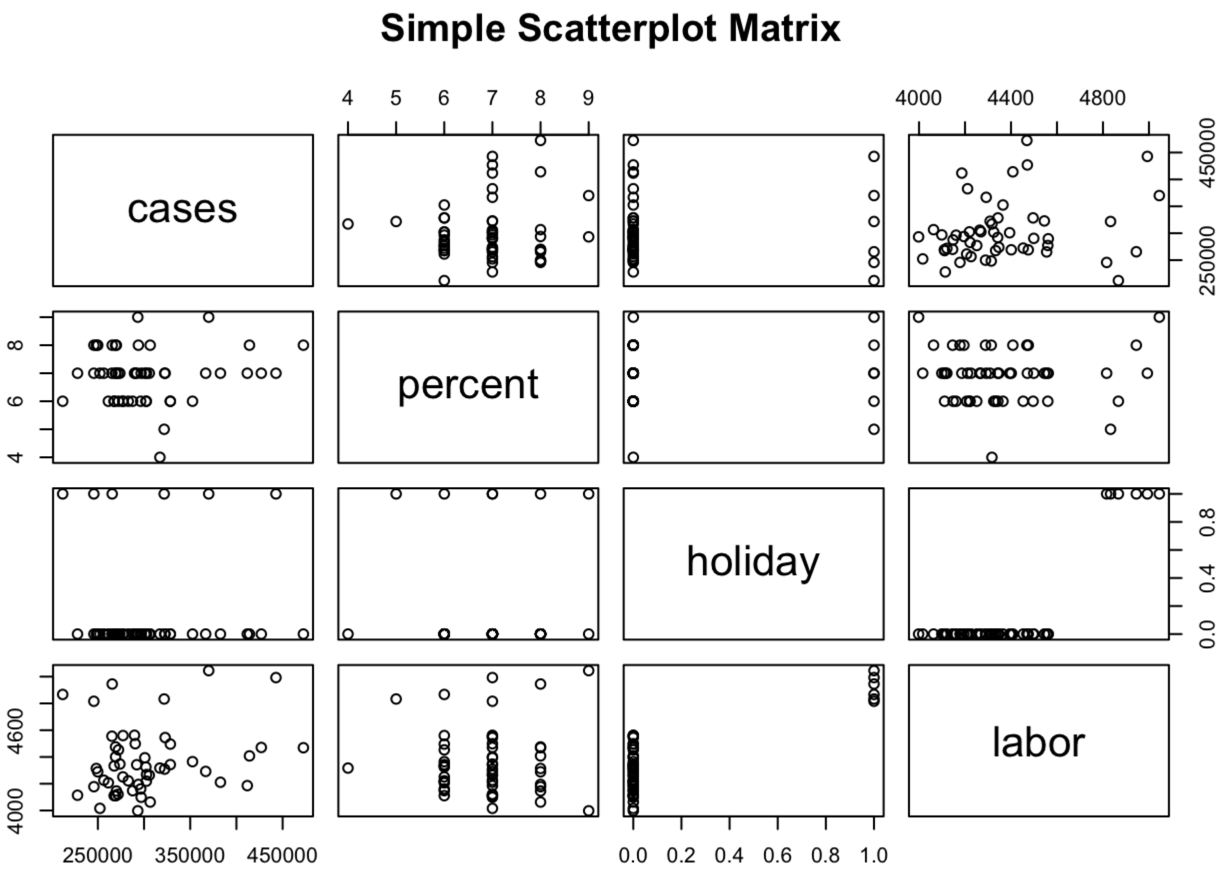


```
plot(time,dat$X3,xlab="time (weeks)",ylab="X3")
```



1. X_1 may depent on time. it has a tendency to be larger over time
2. X_2 is independent of time
3. X_3 seems independent of time

```
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 is obvious that X_3 and Y has a very strong correlation(it also makes sense). the others have no significant correlation.

2

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

```
##
## Call:
## lm(formula = labor ~ cases + percent + holiday)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -259.11 -108.25  -21.61   79.82  294.74
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.139e+03  1.761e+02  23.502  < 2e-16 ***
## cases        7.917e-04  3.651e-04   2.169   0.0351 *
## percent     -1.266e+01  2.141e+01  -0.591   0.5572
## holiday      6.211e+02  6.230e+01   9.970 2.79e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 143.3 on 48 degrees of freedom
## Multiple R-squared:  0.6885, Adjusted R-squared:  0.669
## F-statistic: 35.36 on 3 and 48 DF,  p-value: 3.276e-12
```

```
resid1<-dat.fit$residuals
resid1
```

```
##          1          2          3          4          5          6
##  -28.25828  173.02095  -22.33343  -48.62147   76.06517   22.98588
##          7          8          9         10         11         12
## -153.50148 -163.80825 -136.54499  277.59780  137.61332  -31.33095
##         13         14         15         16         17         18
##  -64.89033 -217.38070   20.07716 -118.43683  174.48028  -75.27944
##         19         20         21         22         23         24
##  105.53395 -186.30323  -49.87943   15.16948 -116.77431   79.78394
##         25         26         27         28         29         30
##   53.54414  -20.88557   79.92480  -54.16952   58.15850  -27.34559
##         31         32         33         34         35         36
## -144.49655 -259.10585  224.32219  239.16016 -233.95633  -62.84747
##         37         38         39         40         41         42
## -142.65327  282.40174 -105.41409  294.74052   23.02814   82.74767
##         43         44         45         46         47         48
##  106.01083   57.32666   42.54117  -83.41947 -142.24191  -28.92922
##         49         50         51         52
##    3.56313  218.77709 -190.25179   60.48504
```

a. The regression function is

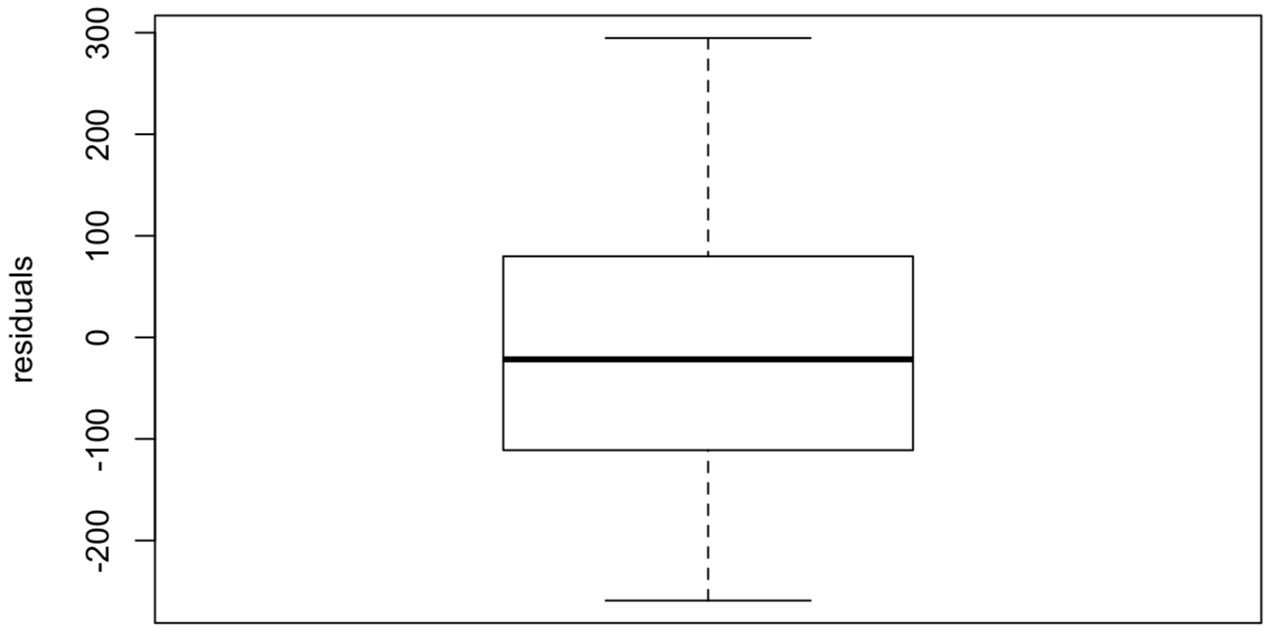
$$Y = 0.0007871X_1 - 13.17X_2 + 623.6X3 + 4150$$

b_1, b_2, b_3 are unbiased estimates of

$$\beta_1, \beta_2, \beta_3$$

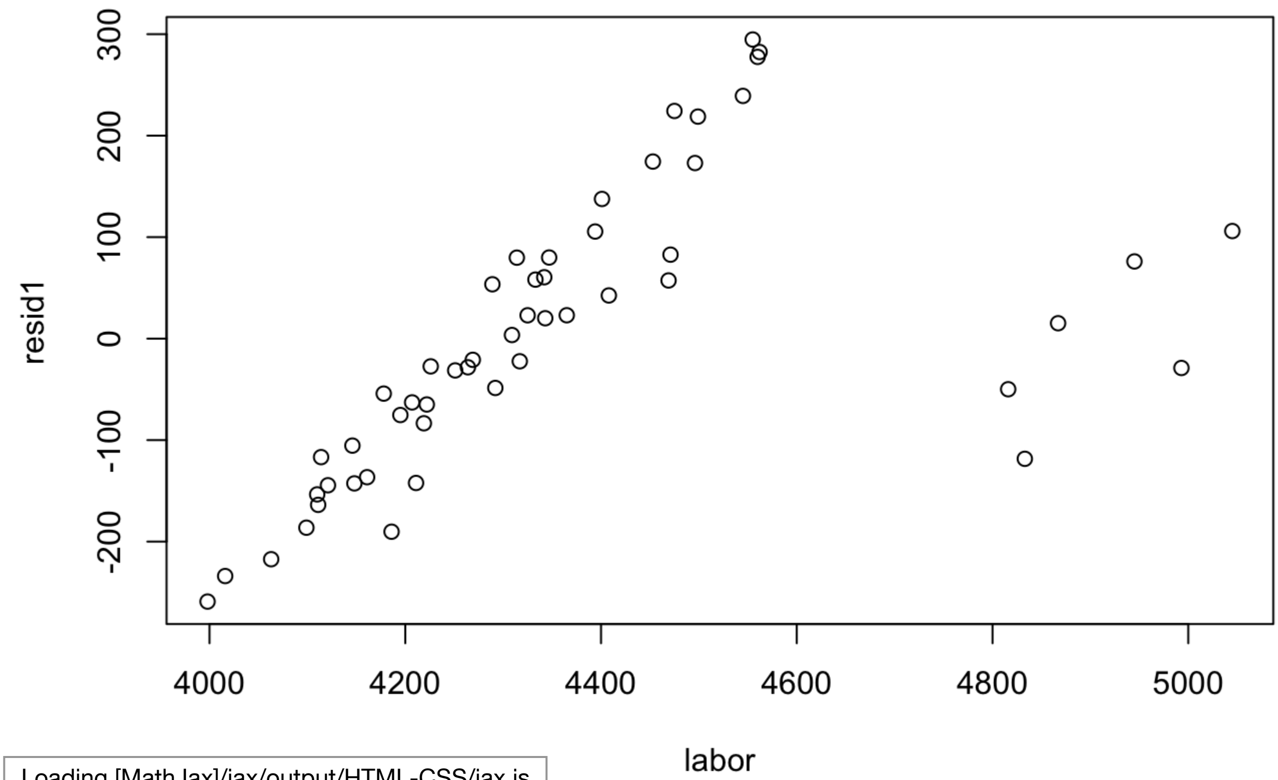
b.

```
boxplot(resid1,ylab="residuals", pch=19)
```



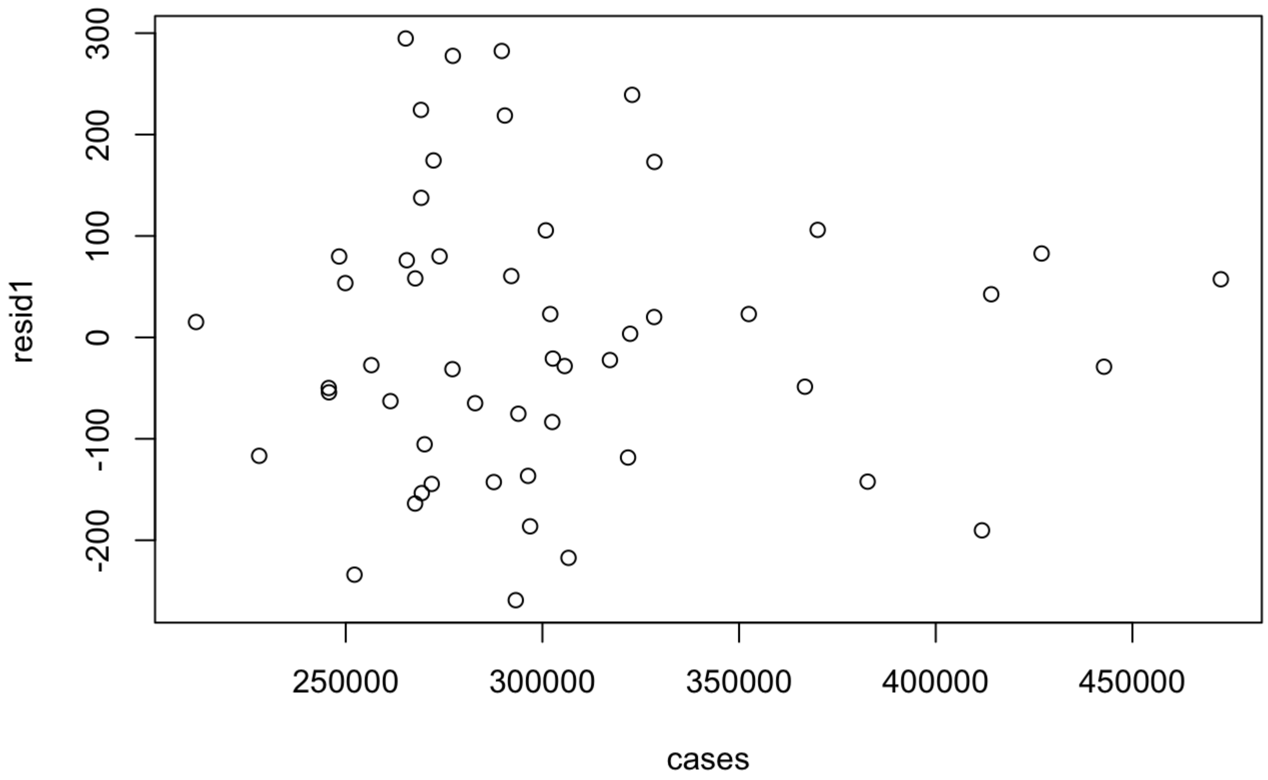
From the boxplot, we can know the median, maximum, minimum, 25 and 75 percent quantile of the residuals.

```
plot(labor,resid1)
```

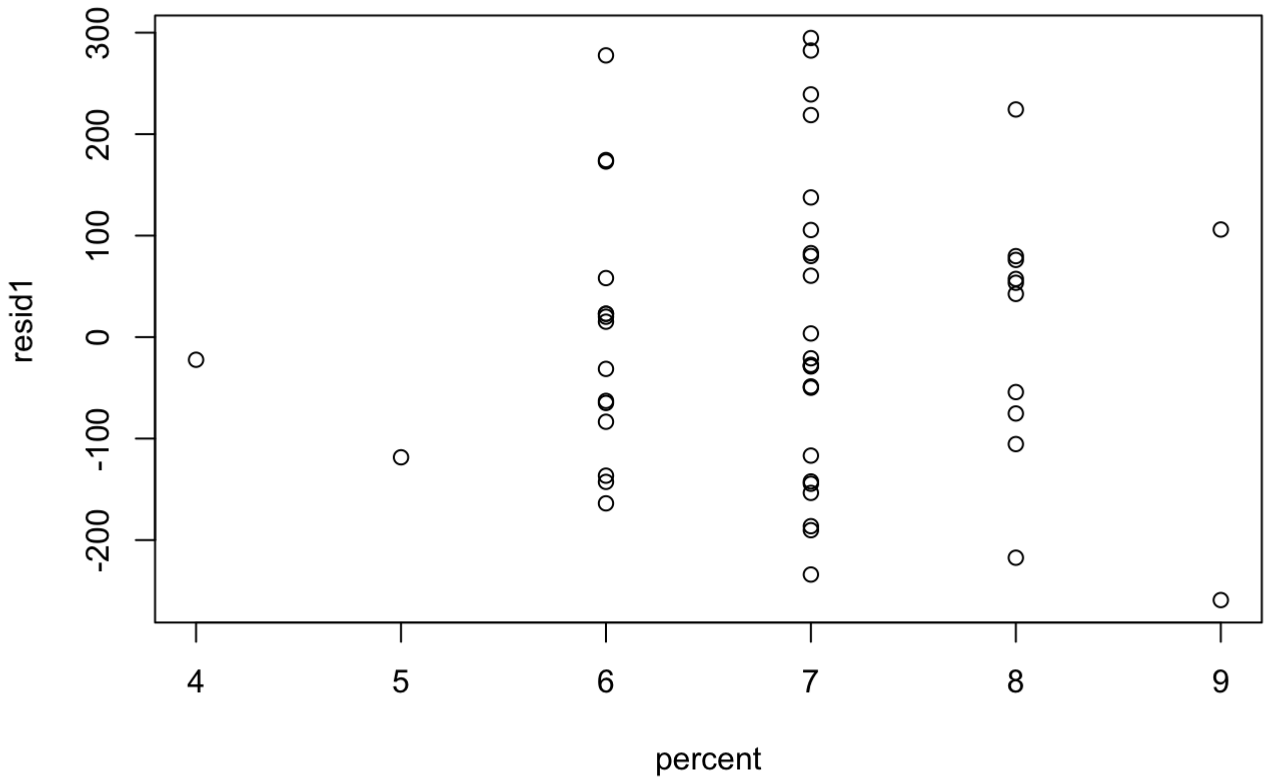


Loading [MathJax]/jax/output/HTML-CSS/jax.js

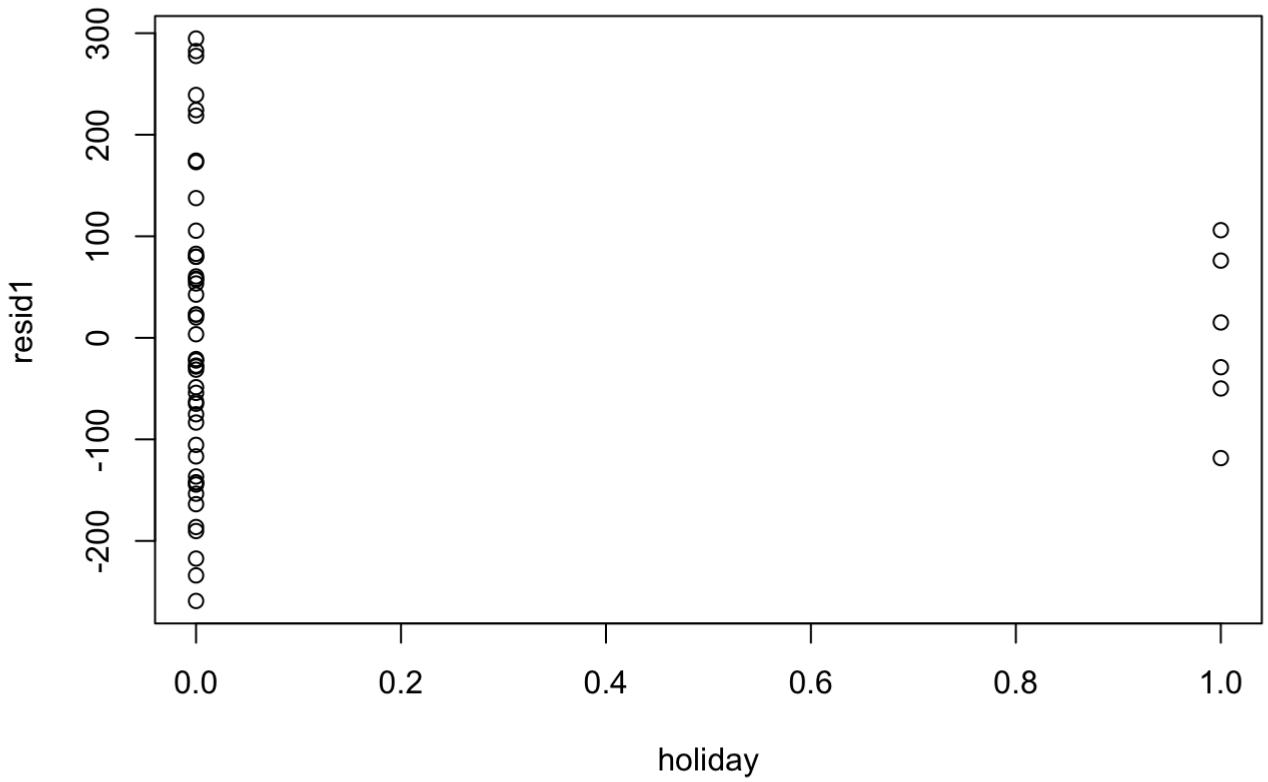
```
plot(cases,resid1)
```



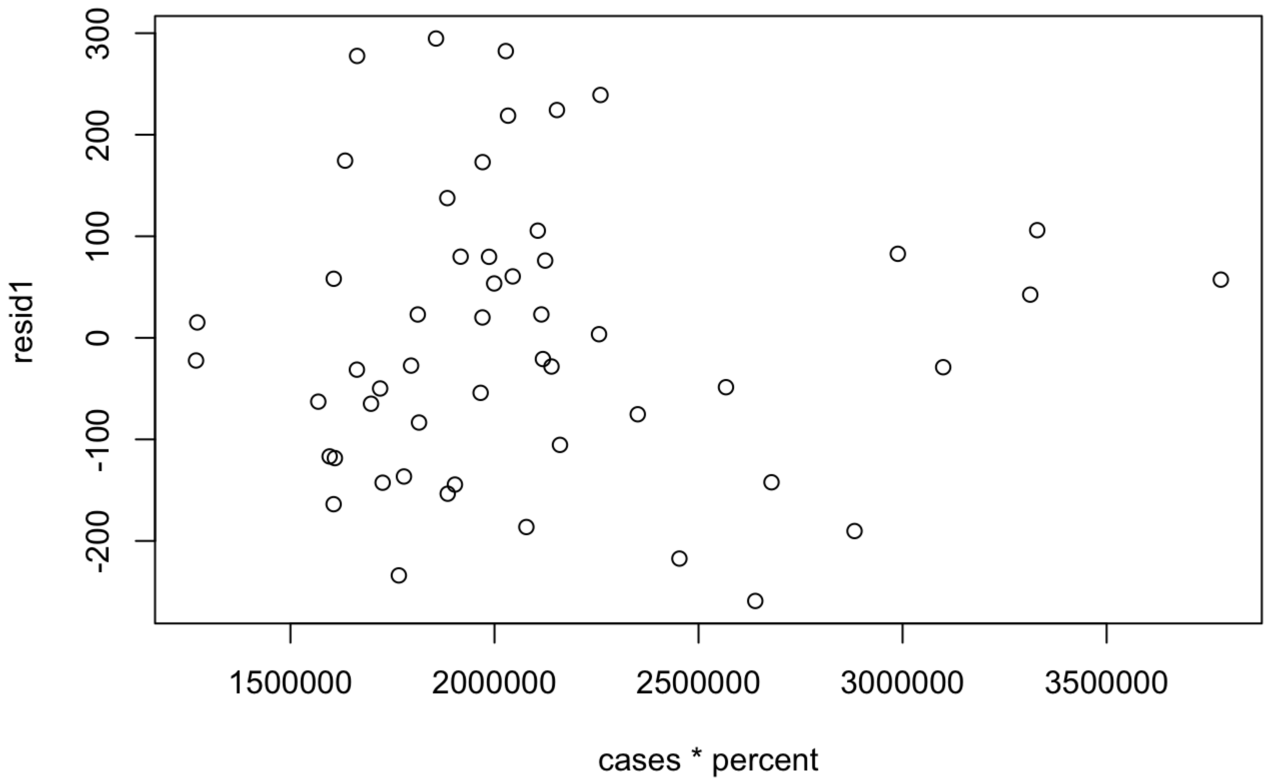

```
plot(percent,resid1)
```



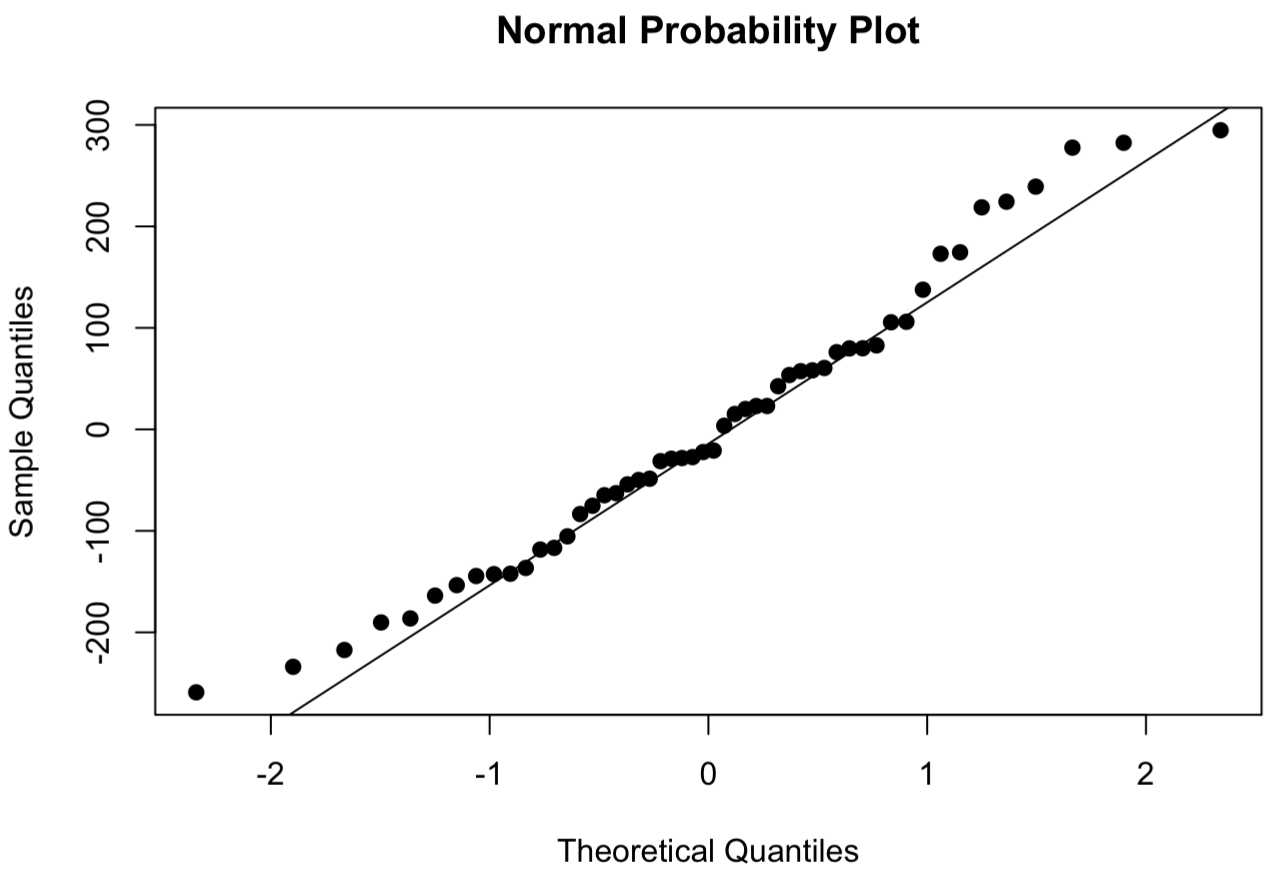

```
plot(holiday,resid1)
```



```
plot(cases*percent,resid1)
```

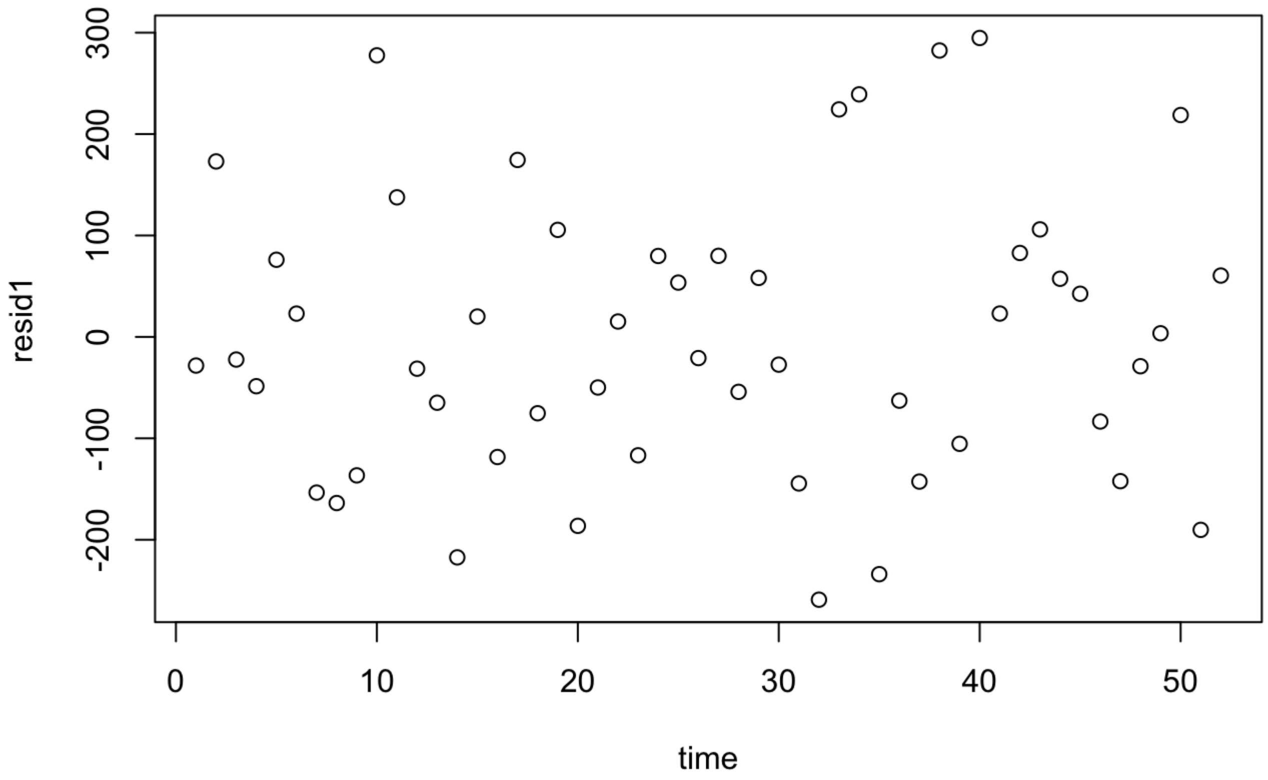



```
qqnorm(resid1, main="Normal Probability Plot", pch=19)
qqline(resid1)
```



c. The plots show that the regression function may not be linear. The residuals change systematically as Y increases, as shown in the first plot. Also the normal probability plot shows that the residuals may not be strictly normally distributed.

```
plot(time,resid1)
```



d)

There does not seem to be any indication that the error terms are correlated.

3

a.

$$H_0:\ \beta_1 = \beta_2 = \beta_3 = 0 \ \ H_a:\ otherwise$$

$$We\ reject\ H_0\ if\ F^* = \frac{MSR}{MSE} > F_{0.95\,,3\,,48}$$

Based on the result:“F-statistic: 35.34 on 3 and 48 DF, p-value: 3.316e-12”, we reject H0 and conclude Ha. The p value is 3.316e-12 The t-test result from above implies that

$$\beta_1\ and\ \beta_3$$

are likely to be non-zero but β_2 may be zero.

b.

```
confint(dat.fit,c(2,4),level = 1-0.05/4)
```

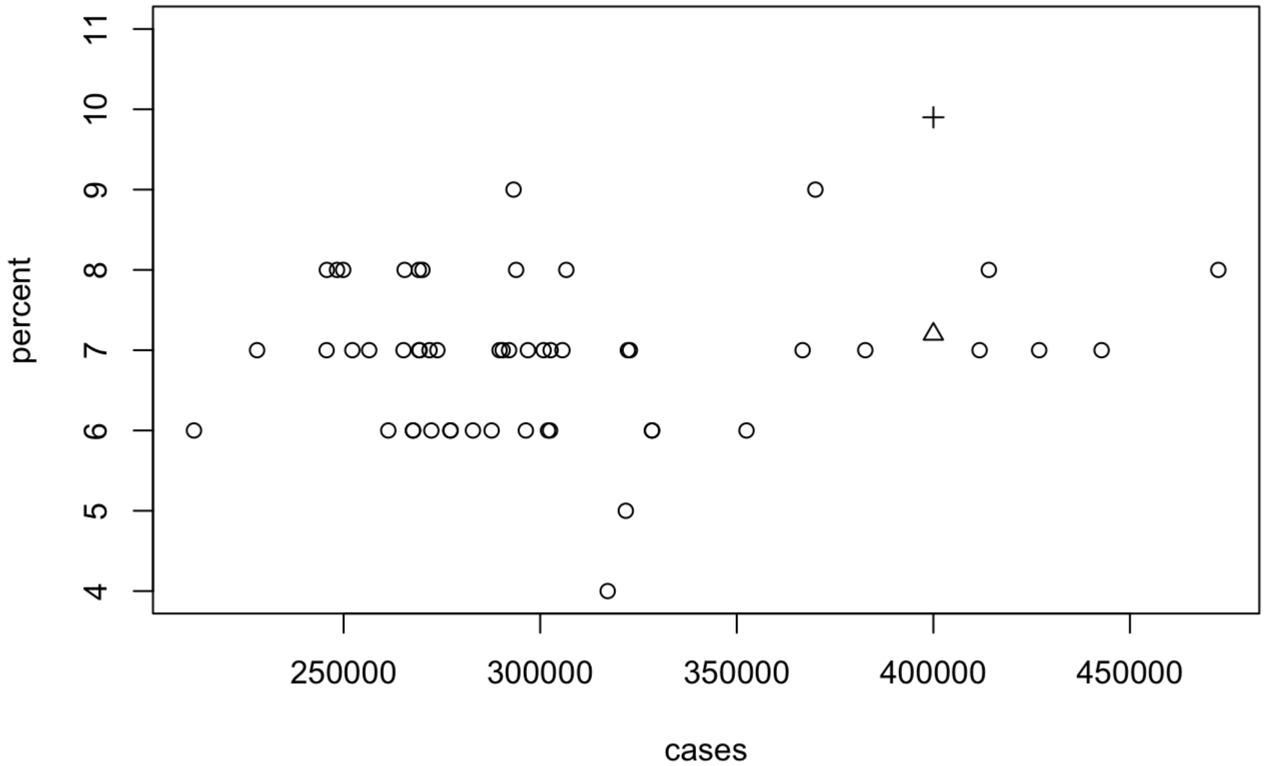
```
##                0.625 %     99.375 %
## cases   -1.557754e-04 1.739169e-03
## holiday  4.594321e+02 7.827869e+02
```

The family confidence interval is shown above. The family confidence coefficient means that when doing many simulations, the proportion of samples which values fall correctly in the cofifence interval.

c.Coefficient of multiple determination is 0.6883. It can be viewed as a coefficient of simple determination between the responses and the fitted values.

4

```
plot(cases, percent,ylim=c(4,11))
points(400000,7.2,pch=2)
points(400000,9.9,pch=3)
```



It is a plot of the two variables: X1 and X2. The cross and triangle represent the two points where predictions are to be made. It can be seen that the triangle lies well within the joint range of the two variables, but the cross seems to be out of the scope of the model.

5

```
new1 <- data.frame(cases=230000,percent=7.5,holiday=0)
new2 <- data.frame(cases=250000,percent=7.3,holiday=0)
new3 <- data.frame(cases=280000,percent=7.1,holiday=0)
new4 <- data.frame(cases=340000,percent=6.9,holiday=0)
predict(dat.fit, new1, se.fit = F, interval = "prediction", level = 1-0.05/8)
```

```
##        fit      lwr      upr
## 1 4226.033 3802.696 4649.371
```

```
predict(dat.fit, new2, se.fit = F, interval = "prediction", level = 1-0.05/8)
```

```
##        fit     lwr      upr
## 1 4244.398 3825.55 4663.247
```

```
predict(dat.fit, new3, se.fit = F, interval = "prediction", level = 1-0.05/8)
```

```
##       fit      lwr      upr
## 1 4270.68 3855.573 4685.787
```

```
predict(dat.fit, new4, se.fit = F, interval = "prediction", level = 1-0.05/8)
```

```
##        fit      lwr      upr
## 1 4320.713 3904.643 4736.783
```

The intervals are presented above.

6

```
new <- data.frame(cases=282000,percent=7.1,holiday=0)
predict(dat.fit, new, se.fit = T, interval = "prediction", level = 1-0.05)
```

```
## $fit
##        fit     lwr      upr
## 1 4272.264 3980.54 4563.987
##
## $se.fit
## [1] 23.01511
##
## $df
## [1] 48
##
## $residual.scale
## [1] 143.2534
```

```
mse <- mean(dat.fit$residuals^2)
mse
```

```
## [1] 18942.95
```

We obtained MSE and se.fit.

```
lwr<-4278.365-qt(1-0.05/2,df=48)*sqrt(mse/3+22.83758^2)
upr<-4278.365+qt(1-0.05/2,df=48)*sqrt(mse/3+22.83758^2)
lwr
```

```
## [1] 4112.127
```

```
upr
```

```
## [1] 4444.603
```

a. The interval is (4112.088,4444.642)
b. Just multiply the interval by 3. We obtain (12336.27,13333.92)