Homework 3

Charles Hwang

10/13/2022

Charles Hwang

Dr. Xi

STAT 408-001

2022 October 13

Problem 1

```
rm(list=ls())
p<-read.csv("/Users/newuser/Desktop/Notes/Graduate/STAT 408 - Applied Regression Analysis/prostate.csv"
l<-summary(lm(lpsa~.,data=p))
t<-c()
```

Problem 1a

```
l$coefficients["age","Estimate"]+c(1,-1)*qnorm(0.05/2)*l$coefficients["age","Std. Error"]
```

```
## [1] -0.041535314  0.002260963
```

We are **95** percent confident β_{age} is between -0.0415353 and 0.002261.

Problem 1b

```
l$coefficients["age","Estimate"]+c(1,-1)*qnorm(0.1/2)*l$coefficients["age","Std. Error"]
```

```
## [1] -0.038014673 -0.001259678
```

We are **90** percent confident β_{age} is between -0.0380147 and -0.0012597.

Problem 1c

```
l$coefficients["age","Pr(>|t|)"]
```

```
## [1] 0.08229321
```

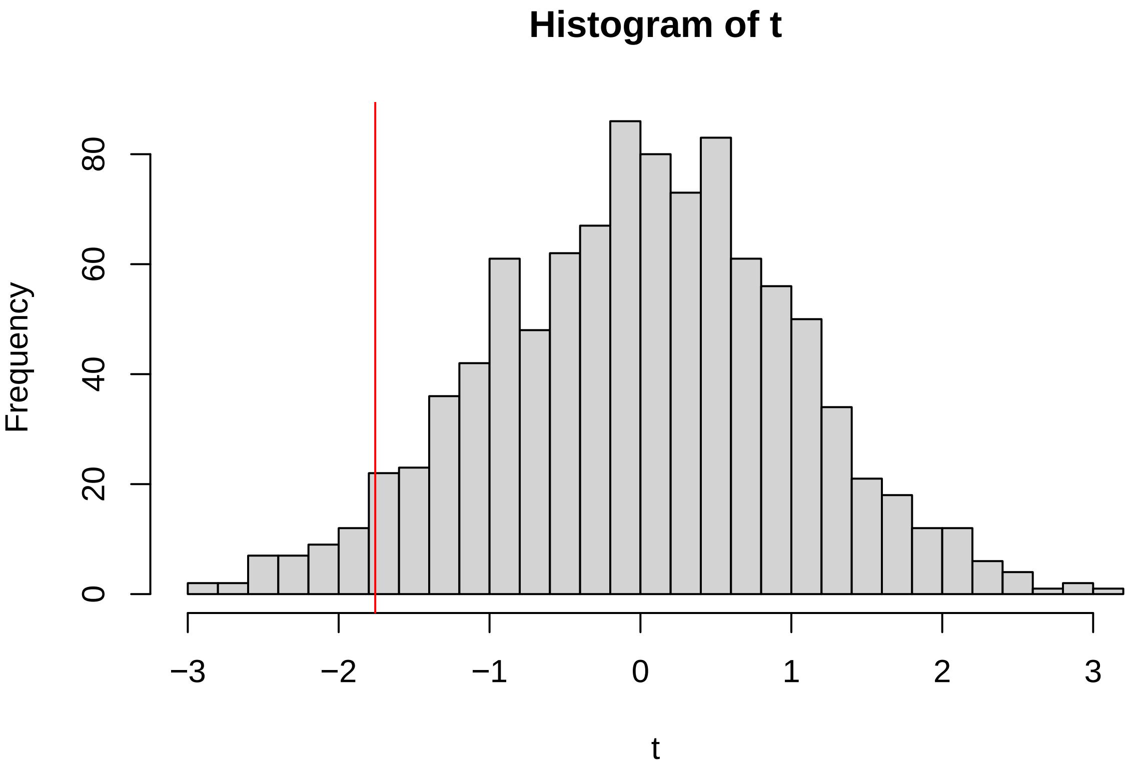
Based on the confidence intervals found in problems 1a and 1b, we can expect $0.05 < p < 0.1$. This is consistent with the summary output showing $p = 0.0822932$.

Problem 1d

```
set.seed(1610)
for(i in 1:1000){r<-lm(lpsa~lcavol+lweight+sample(age)+lbph+svi+lcp+gleason+pgg45,data=p)
t[i]<-summary(r)$coefficients["sample(age)","t value"]}
l$coefficients["age","t value"]
```

```
## [1] -1.757599
```

```
hist(t,breaks=25)
abline(v=l$coefficients["age","t value"],col="red")
```



```
mean(abs(t)>abs(l$coefficients["age","t value"]))
```

```
## [1] 0.09
```

The permutation t-test yields a p-value of 0.09. This is very similar to p-value from the summary output printed in problem 1c (0.0822932).

Problem 1e

```
l$coefficients[,"Pr(>|t|)"]<0.05
```

```
## (Intercept)      lcavol     lweight         age        lbph         svi
##       FALSE        TRUE        TRUE       FALSE       FALSE        TRUE
##         lcp     gleason       pgg45
##       FALSE       FALSE       FALSE
```

```
anova(lm(lpsa~lcavol+lweight+svi,data=p),lm(lpsa~.,data=p))
```

```
## Analysis of Variance Table
##
## Model 1: lpsa ~ lcavol + lweight + svi
## Model 2: lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     93 47.785
## 2     88 44.163  5    3.6218 1.4434 0.2167
```

We fail to reject H_0 at the $\alpha = 0.05$ level. There is insufficient evidence ($F = 1.4433869$, $p = 0.2167334$) that the full model is better than the reduced model. We conclude the reduced model is better.

Problem 2

```
c<-read.csv("/Users/newuser/Desktop/Notes/Graduate/STAT 408 - Applied Regression Analysis/cheddar.csv")
```

Problem 2a

```
summary(lm(taste~.,data=c))$coefficients[,"Pr(>|t|)"]
```

```
## (Intercept)      Acetic         H2S      Lactic
## 0.155399149 0.941979774 0.004247081 0.031079481
```

We can see the variables for both hydrogen sulfide ($p = 0.0042471$) and lactic acid ($p = 0.0310795$) are statistically significant at the $\alpha = 0.05$ level, but the variable for acetic acid is not ($p = 0.9419798$).

Problem 2b

```
summary(lm(taste~exp(Acetic)+exp(H2S)+Lactic,data=c))$coefficients[,"Pr(>|t|)"]
```

```
## (Intercept) exp(Acetic)    exp(H2S)      Lactic
##  0.10419810  0.23711453  0.07856795  0.01046242
```

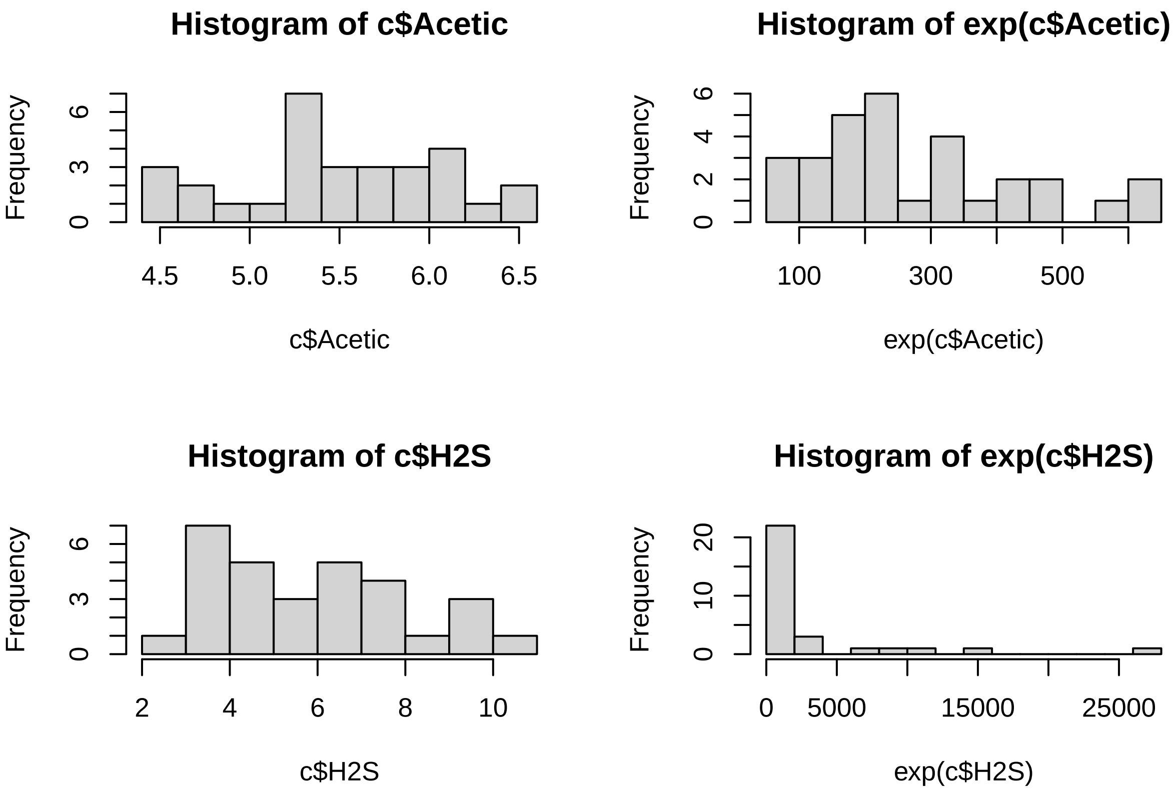
We can see the variable for lactic acid is statistically significant ($p = 0.0104624$) at the $\alpha = 0.05$ level, but the variables for acetic acid ($p = 0.2371145$) and hydrogen sulfide ($p = 0.0785679$) are not.

Problem 2c

```
anova(lm(taste~.,data=c),lm(taste~exp(Acetic)+exp(H2S)+Lactic,data=c))
```

```
## Analysis of Variance Table
##
## Model 1: taste ~ Acetic + H2S + Lactic
## Model 2: taste ~ exp(Acetic) + exp(H2S) + Lactic
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
## 1     26 2668.4
## 2     26 3253.6  0    -585.2
```

```
par(mfrow=c(2,2))
hist(c$Acetic,breaks=10)
hist(exp(c$Acetic),breaks=10)
hist(c$H2S,breaks=10)
hist(exp(c$H2S),breaks=10)
```



We should not use an F-test to compare these two models as they have the same number of degrees of freedom and the difference is 0 degrees of freedom, which is in the denominator of the F-statistic. There is also a negative sum of squares (-585.1972661) which indicates something is clearly violated. However, we can plot histograms of the variables for hydrogen sulfide and lactic acid and their original scale, and we can see the logarithmically transformed versions of the variables are closer to normal. Thus, we can reasonably conclude the original model in problem 2a provides a better fit for these data.

Problem 2d

```
0.01*summary(lm(taste~.,data=c))$coefficients["H2S","Estimate"]
```

```
## [1] 0.03911841
```

We would expect an increase in approximately 0.0391184 average taste score points if hydrogen sulfate were to be increased by 0.01 units, holding all other variables constant.

Problem 3

```
g<-read.csv("/Users/newuser/Desktop/Notes/Graduate/STAT 408 - Applied Regression Analysis/teengamb.csv"
```

Problem 3a

```
summary(lm(gamble~.,data=g))$coefficients[,"Pr(>|t|)"]
```

```
##  (Intercept)          sex       status       income       verbal 
## 1.967736e-01 1.011184e-02 8.534869e-01 1.791882e-05 1.803109e-01
```

We can see the variables for sex ($p = 0.0101118$) and income ($p = 1.7918821 \times 10^{-5}$) are statistically significant at the $\alpha = 0.05$ level, but the variables for socioeconomic status ($p = 0.8534869$) and verbal score ($p =$

0.1803109) are not.

Problem 3b

Yes, the variables that the model found statistically significant make sense. It is reasonable to expect a difference in sex in amount of gambling, and intuitively, income should be a significant variable in the model.

Problem 3c

```
anova(lm(gamble~.,data=g),lm(gamble~income,data=g))
```

```
## Analysis of Variance Table
##
## Model 1: gamble ~ sex + status + income + verbal
## Model 2: gamble ~ income
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1     42 21624
## 2     45 28009 -3   -6384.8 4.1338 0.01177 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

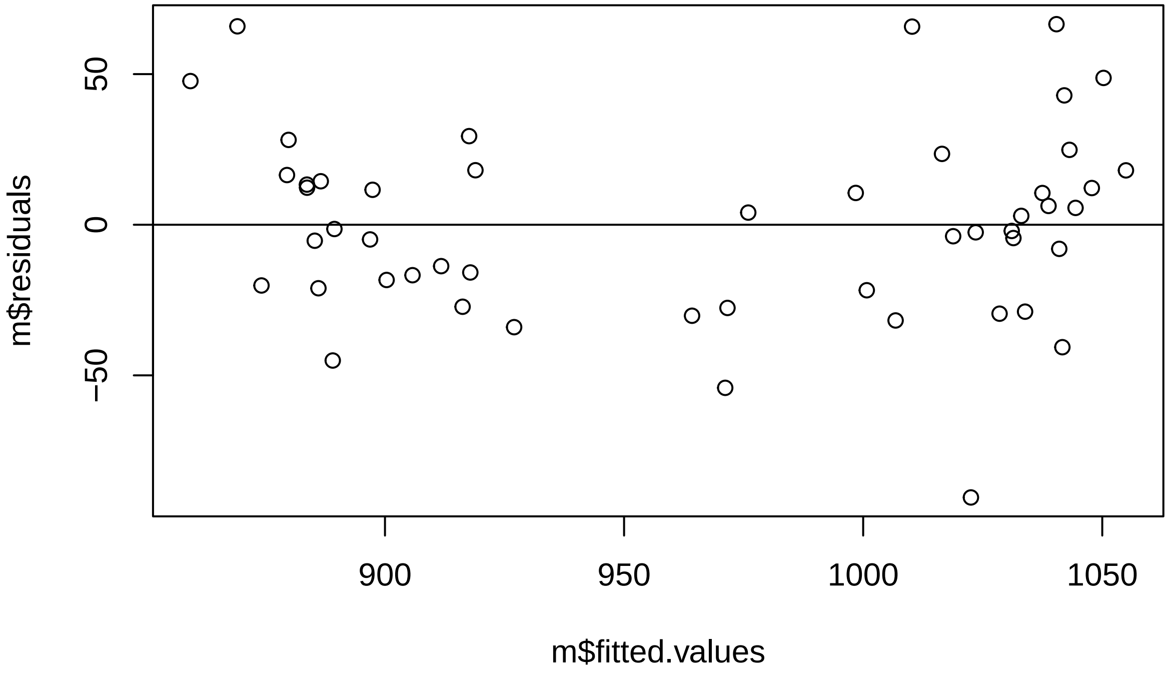
We reject H_0 at the $\alpha = 0.05$ level. There is sufficient evidence ($F = 4.1337611$, $p = 0.0117721$) that the full model is better than the reduced model with only `income` as a predictor variable.

Problem 4

```
s<-read.csv("/Users/newuser/Desktop/Notes/Graduate/STAT 408 - Applied Regression Analysis/sat.csv")
m<-lm(total~expend+salary+ratio+takers,data=s)
```

Problem 4a

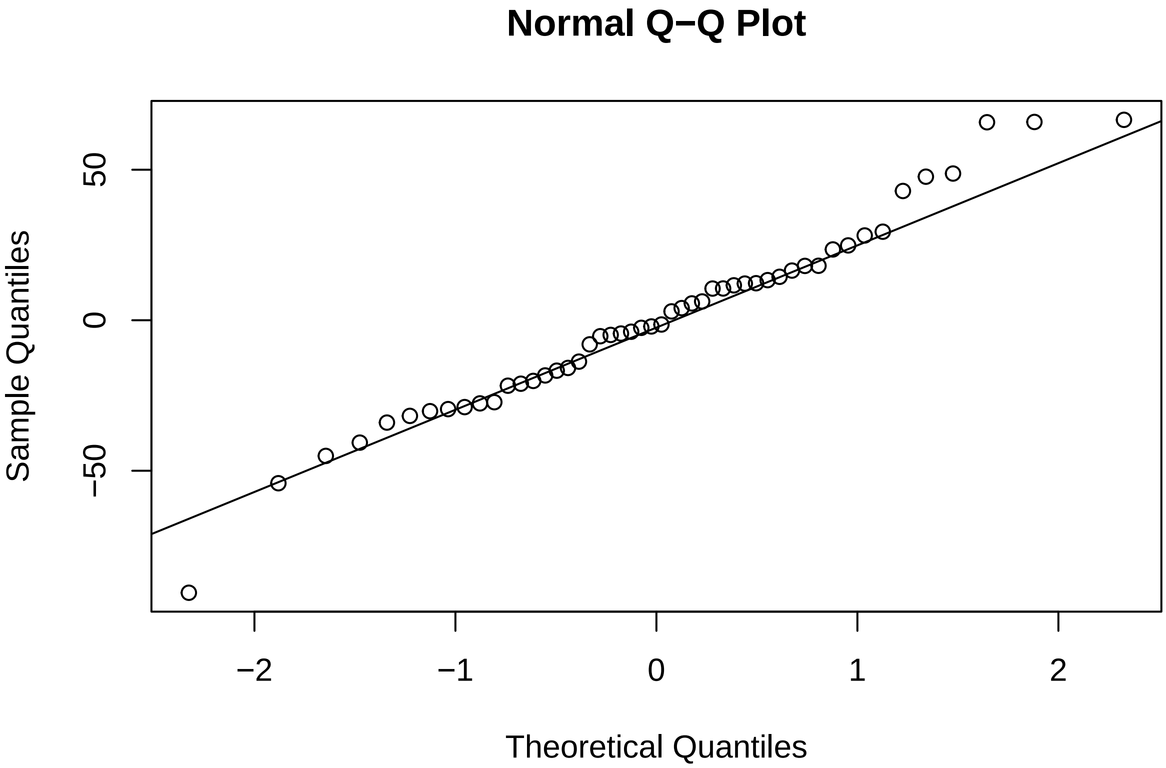
```
plot(m$residuals~m$fitted.values)
abline(h=0)
```



There does not appear to be any pattern in the residuals vs. fitted values plot.

Problem 4b

```
qqnorm(m$residuals)
qqline(m$residuals)
```



There is slight tail at each of the ends of the Q-Q plot, but the normality assumption does not appear to be badly violated.

Problem 4c

```
rstudent(m)
```

```
##           1           2           3           4           5           6 
## -0.06574555 -1.02091600 -0.86394293 -0.90546917 -0.56764583  0.12375609 
##           7           8           9          10          11          12 
##  0.97811087  0.41584816 -0.89619150 -0.64380944 -0.52185887 -0.68138366 
##          13          14          15          16          17          18 
##  0.34240342 -0.57108388  1.56037417  0.38967283 -0.92044224 -0.08013841 
##          19          20          21          22          23          24 
##  0.53951189  0.36099300  1.55127336 -0.26522896  1.36779678  0.09173934 
##          25          26          27          28          29          30 
##  0.19418490  0.34070604  0.17614652 -1.73200396  2.19000605 -0.47250752 
##          31          32          33          34          35          36 
## -0.11823952 -0.16415440 -0.67193751  2.21368580 -0.99430251 -0.13785064 
##          37          38          39          40          41          42 
##  0.94590973 -0.17154457 -0.04457306 -1.46883156  0.79638212  0.75084252 
##          43          44          45          46          47          48 
## -1.07088136  2.52958734  0.45748671  0.40448538  0.58126580 -3.12442832 
##          49          50
##  0.57665563 -1.31188995
```

```
sum(abs(rstudent(m))>qt(1-0.05/2,abs(diff(dim(s)))-1))
```

```
## [1] 4
```

```
rstudent(m)[which(abs(rstudent(m))>qt(1-0.05/2,abs(diff(dim(s)))-1))]
```

```
##        29        34        44        48
##  2.190006  2.213686  2.529587 -3.124428
```

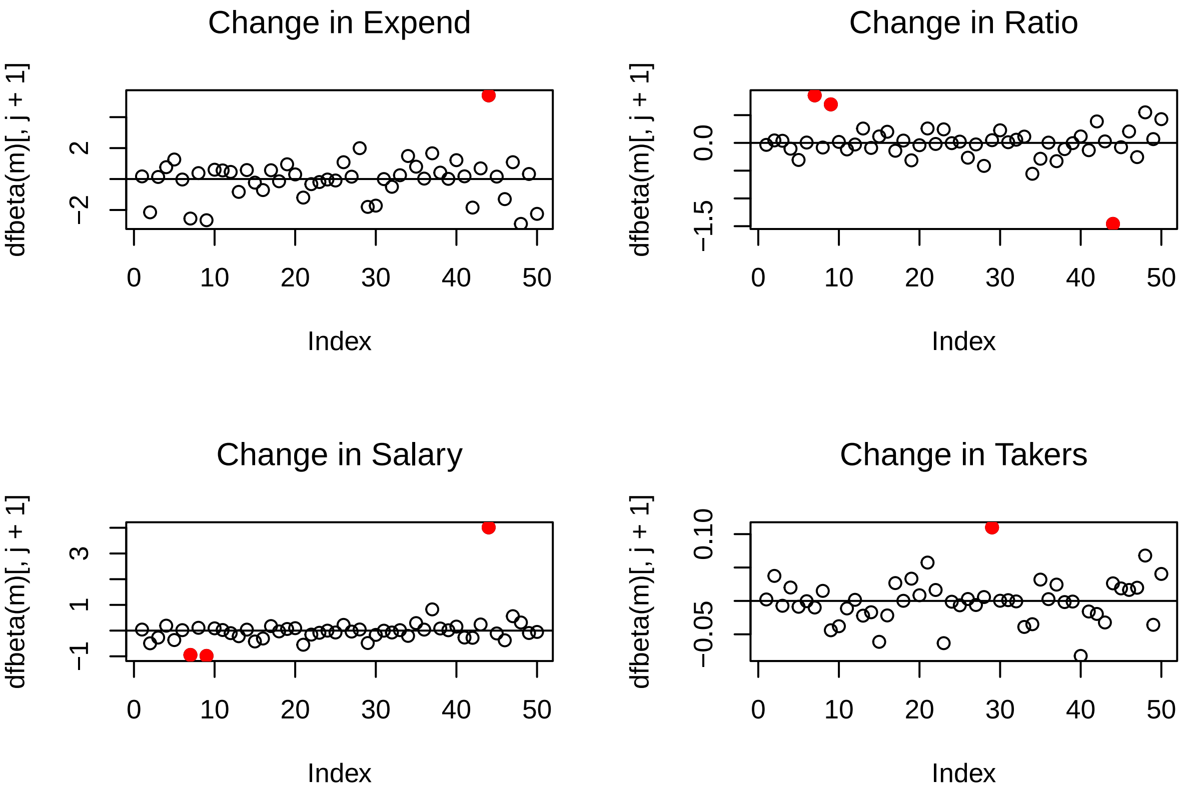
We can see that observations 29, 34, 44, 48 appear to be statistical outliers. (Lecture 9, Slide 21)

Problem 4d

```
dfbeta(m)
```

```
##      (Intercept)       expend       salary        ratio        takers
## 1   -0.586403914  0.171524024 -0.035367792  0.040519463  0.0021718503
## 2   17.396120028 -2.149510788  0.044970559 -0.493880467  0.0374488215
## 3    2.118317370  0.133643810  0.038103401 -0.270394241 -0.0073063709
## 4   -5.430112877  0.765318960 -0.104416098  0.191601957  0.0200718436
## 5    9.319925361  1.262393355 -0.307382495 -0.367828993 -0.0087839101
## 6   -0.249380773 -0.030691226  0.006629759  0.017617858 -0.0004025201
## 7    2.290412509 -2.552166802  0.856096305 -0.946789147 -0.0098724016
## 8   -1.435223820  0.383620590 -0.083479172  0.108616045  0.0150159148
## 9    9.064963850 -2.660731688  0.693599049 -0.984377602 -0.0439561726
## 10  -4.778177105  0.601989897  0.017927106  0.088354063 -0.0379091071
## 11   0.469306098  0.556788577 -0.117440493  0.022360511 -0.0112732858
## 12  -0.555908524  0.459798583 -0.028847471 -0.099256107  0.0016067486
## 13   0.578807620 -0.829609626  0.259170491 -0.218557724 -0.0220989693
## 14  -0.692279868  0.580099560 -0.090374473  0.037169415 -0.0169834606
## 15   7.624373899 -0.232017774  0.117008586 -0.423226633 -0.0612712658
## 16   3.435607517 -0.712500840  0.201037778 -0.307985445 -0.0217428937
## 17  -2.943736945  0.567004443 -0.141426574  0.175861794  0.0266934773
## 18  -0.136440208 -0.145804560  0.042100938 -0.031403865  0.0001431234
## 19   3.476596149  0.952606616 -0.314222020  0.062402043  0.0333010614
## 20  -1.959035495  0.298247400 -0.041605960  0.094576936  0.0084453132
## 21   6.321680476 -1.198204248  0.259864989 -0.550051613  0.0573945294
## 22   4.467009025 -0.329259379 -0.018991020 -0.155962876  0.0163490491
## 23  -2.776002206 -0.193538009  0.243945490 -0.085030697 -0.0632739721
## 24   0.498622146 -0.032445154 -0.004672445 -0.002309139 -0.0012116810
## 25   1.347845746 -0.094405456  0.023207002 -0.073217202 -0.0065892563
## 26  -0.738069855  1.086334447 -0.266481696  0.222101897  0.0028413902
## 27   1.027887684  0.152675726 -0.027213801 -0.038257716 -0.0061097101
## 28   0.582304053  1.994665866 -0.414202557  0.043135317  0.0057674373
## 29  14.646103897 -1.800348133  0.048645150 -0.483543339  0.1100742451
## 30   4.698895535 -1.718561923  0.227101123 -0.167400259  0.0002427227
## 31  -0.467662594 -0.003675247  0.012219580 -0.003349841  0.0011411132
## 32   2.108202039 -0.500904695  0.057314143 -0.073646653 -0.0007265971
## 33  -4.786596302  0.245057697  0.113501306  0.017569139 -0.0389764790
## 34  16.729964836  1.488615402 -0.556229814 -0.205788968 -0.0348950642
## 35  -1.467852637  0.798173808 -0.285695199  0.291575089  0.0318968842
## 36  -1.177429067  0.027908651  0.005156958  0.038284478  0.0026564217
## 37 -12.490762254  1.663743735 -0.329112418  0.825778386  0.0243999076
## 38   0.005947482  0.413232385 -0.111508101  0.081695506 -0.0017256698
## 39  -0.117138819  0.010675483 -0.004663351  0.013039851 -0.0009601844
## 40 -11.994235592  1.218489573  0.117005958  0.155848214 -0.0823435130
## 41   9.207398751  0.174757704 -0.131490475 -0.269964011 -0.0158462537
## 42   3.348353360 -1.844896005  0.386969132 -0.280348161 -0.0195119994
## 43  -8.612239141  0.689200936  0.027214249  0.238043098 -0.0322760092
## 44 -47.874437434  5.405333540 -1.458512248  4.014911955  0.0263238657
## 45   3.300667172  0.157485942 -0.078165489 -0.109837853  0.0187033768
## 46   6.564763637 -1.297545365  0.205991068 -0.377627062  0.0164362079
## 47  -7.246213545  1.087092733 -0.255961216  0.560670514  0.0197298231
## 48 -11.705969699 -2.896706205  0.550385435  0.315502872  0.0679129776
## 49  -1.064528392  0.326328908  0.068082322 -0.093231925 -0.0358100792
## 50  -3.049577948 -2.249554911  0.427828307 -0.052321248  0.0404061712
```

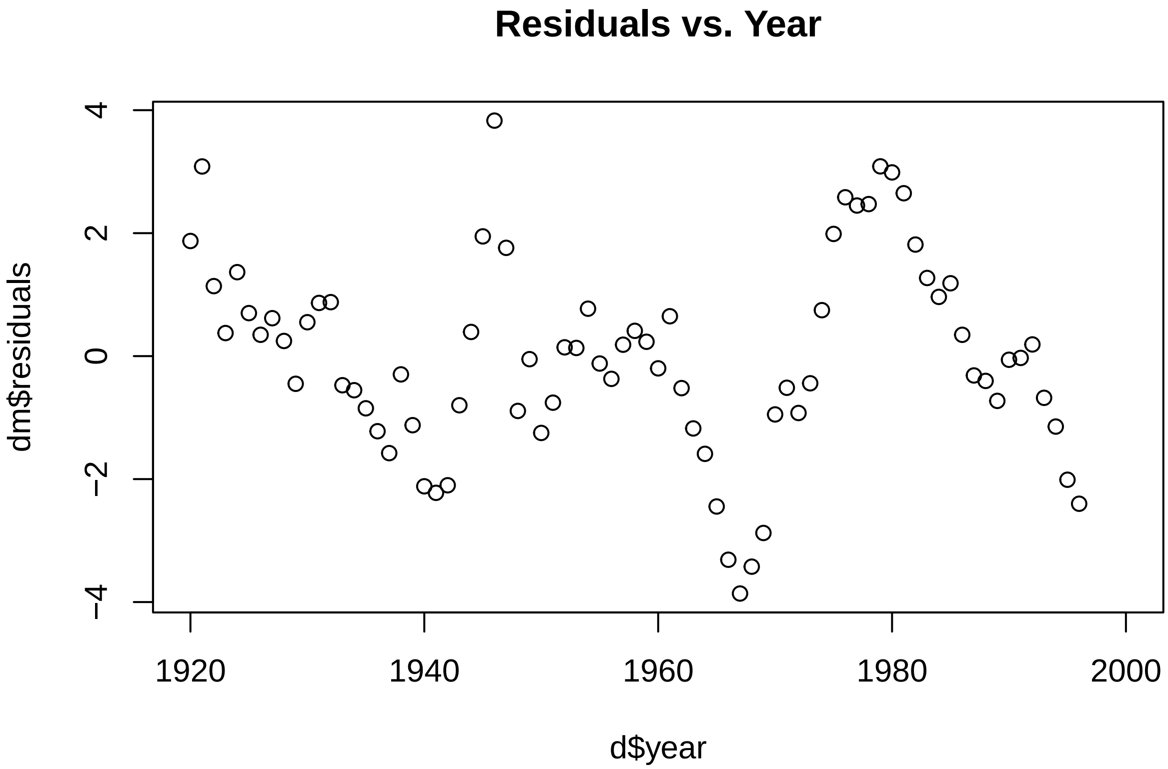
```
library(tools)
ts<-toTitleCase(colnames(s))[c(1:4)]
par(mfrow=c(2,2))
for(j in 1:4){plot(dfbeta(m)[,j+1],main=substitute(paste("Change in ",x),list(x=ts[j])))
  e<-which(abs(dfbeta(m)[,j+1])>3*IQR(dfbeta(m)[,j+1]))
  points(e,dfbeta(m)[e,j+1],pch=19,col="red")
  abline(h=0)}
```



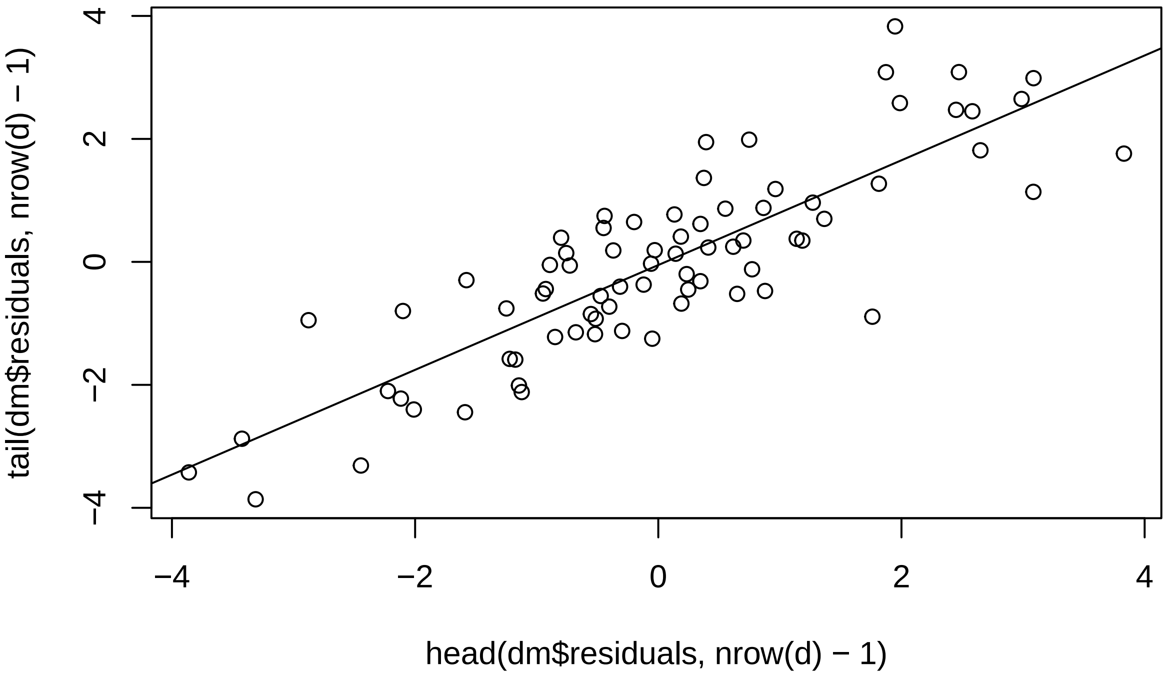
We can see that observation 44 is an influential point for the expend variable; observations 7, 9, 44 are influential points for the salary variable; observations 7, 9, 44 are influential points for the ratio variable; and observation 29 is an influential point for the takers variable.

Problem 5

```
d<-read.csv("/Users/newuser/Desktop/Notes/Graduate/STAT 408 - Applied Regression Analysis/divusa.csv")
dm<-lm(divorce~unemployed+femlab+marriage+birth+military,data=d)
plot(dm$residuals~d$year,main="Residuals vs. Year",xlim=round(range(d$year),-1))
```



```
plot(tail(dm$residuals,nrow(d)-1)~head(dm$residuals,nrow(d)-1))
ds<-summary(lm(tail(dm$residuals,nrow(d)-1)~head(dm$residuals,nrow(d)-1)))
abline(ds$coefficients["(Intercept)","Estimate"],ds$coefficients["head(dm$residuals, nrow(d) - 1)","Esti
```



```
ds
```

```
##
## Call:
## lm(formula = tail(dm$residuals, nrow(d) - 1) ~ head(dm$residuals,
##     nrow(d) - 1))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.34046 -0.54703 -0.08307  0.47315  2.22203
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -0.05155    0.09769  -0.528    0.599
## head(dm$residuals, nrow(d) - 1)  0.85213    0.06218  13.705   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8515 on 74 degrees of freedom
## Multiple R-squared:  0.7174, Adjusted R-squared:  0.7136
## F-statistic: 187.8 on 1 and 74 DF,  p-value: < 2.2e-16
```

We can see from the first plot that the data are correlated by year as there is a clear sinusoidal pattern. The second plot shows no apparent pattern in the residuals of observation $t + 1$ vs. the residuals of observation t.