Final

Aurora Travers

3/7/2022

Q1)

```
knitr::opts_chunk$set(error = TRUE)
dat = read.delim2('Demographic.txt',header=F,sep ='')
names(dat) = c('id','county','state','land_area','total_population','percent_population_18to34','percen
               'No_physicians','No_hospital_beds','crimes','highschool','bachelor',
               'below_poverty','unemployment','per_cap_income','total_income','geo_region')

regional_data_generator2 = function(dat,region){
  loader = dat[dat$geo_region == region,]
  n = nrow(loader)
  y = matrix(loader$total_income,ncol=1,nrow=n)
  x1 = as.numeric(loader$bachelor)
  x2 = as.numeric(loader$highschool)
  x3 = as.numeric(loader$percent_population_18to34)
  x4 = as.numeric(loader$percent_population_65orOlder)
  x5 = as.numeric(loader$unemployment)
  x = cbind(rep(1,n),x1,x2,x3,x4,x5)
  x = as.matrix(x)
  out = list('y' =y,'x' =x)
  return(out)
}

loader1 =regional_data_generator2(dat= dat,region =1)
loader2 =regional_data_generator2(dat= dat,region =2)
loader3 =regional_data_generator2(dat= dat,region =3)
loader4 =regional_data_generator2(dat= dat,region =4)

#how to write your own lm() function
MT2_regression_model = function(loader,alpha){
  #generate matrix
  y = loader$y
  x = loader$x
  #get dimension
  n = nrow(y)
  p = ncol(x)
  #b
  b = solve(t(x)%*%x) %*% t(x) %*% y
  #create matrix
  J = matrix(1,nrow = n,ncol = n)
  I = diag(rep(1,n))
  H = x %*% solve(t(x)%*%x) %*% t(x)
```

```r
  #residuals
  e = (I-H) %*% y
  #quadratic form
  SSE = c(t(y) %*% (I - H) %*% y)
  SSR = c(t(y) %*% (H-(1/n)*J) %*% y)
  #MSE and MSR
  MSE = SSE/(n-p)
  MSR = SSR/(p-1)
  #F statistic
  F_obs = MSR/MSE
  F_critical = qf(p=1-alpha,lower.tail = T,df1 = p-1,df2 = n-p)
  #test result
  if(F_obs>F_critical){Ftest_result = paste('Reject the null when alpha equals to',alpha)}
  if(F_obs<F_critical){Ftest_result = paste('Fail to reject the null when alpha equals to',alpha)}

  return(list('b'=b,'F_obs' = F_obs, 'F_critical_value' = F_critical, 'Test_result' = Ftest_result,'res

}
test_run1 = MT2_regression_model(loader=loader1,alpha=0.01)
test_run2 = MT2_regression_model(loader=loader2,alpha=0.01)
test_run3 = MT2_regression_model(loader=loader3,alpha=0.01)
test_run4 = MT2_regression_model(loader=loader4,alpha=0.01)
```

model 1: Y hat: 63937.2532 + 1224.7094X1 - 938.2231X2 - 509.4389X3 + 260.5753X4 + 227.2832X5 MSE: 52134954 MSR: 535539815 p value = 3.211911 Conclusion: reject the null when alpha equals to 0.01

In region 1, b2 and b3 are negative. This means as more high school graduates and percent population 18 to 34 is present, there should be less serious crimes.

model 2: Y hat: 151391.7382 + 1652.0451X1 - 1774.6043X2 - 1012.1532X3 - 849.5406X4 + 379.4847X5 MSE: 114628083 MSR: 767567960 p value = 3.20205 Conclusion: reject the null when alpha equals to 0.01

In region 2, b2, b3, and b4 are negative. This means as more high school graduates, percent population 18 to 34, and percent population 65 or older is present, there should be less serious crimes.

model 3: Y hat: 3001.242075 + 410.052422X1 - 27.342475X2 - 213.958514X3 + 1.393427X4 + 452.223413X5 MSE: 49737761 MSR: 205419970 p value = 3.145066 Conclusion: reject the null when alpha equals to 0.01

In region 3, b2 is negative. This means as more high school graduates, there should be less serious crimes.

model 4: Y hat: 138568.3499 + 1007.2152X1 - 1809.1304X2 + 798.7147X3 - 885.9512X4 - 2494.6207X5 MSE: 466772558 MSR: 1292140764 p value = 3.286652 Conclusion: fail to reject the null when alpha equals to 0.01

In region 4, b2, b4, and b5 are negative. This means as more high school graduates, percent population 65 or older, unemployment is present, there should be less serious crimes.

Decision Rule: If F* <= F(1-a, p-1, n-p), conclude H0 If F* > F(1-a, p-1, n-p), conclude Ha

```r
test_run1$b
```

```
##            [,1]
##      63937.2532
## x1    1224.7094
## x2    -938.2231
## x3    -509.4389
## x4     260.5753
## x5     227.2832
```

```r
test_run2$b
```

```
##           [,1]
##    151391.7382
## x1   1652.0451
## x2  -1774.6043
## x3  -1012.1532
## x4   -849.5406
## x5    379.4847
```

```r
test_run3$b
```

```
##           [,1]
##    3001.242075
## x1  410.052422
## x2  -27.342475
## x3 -213.958514
## x4    1.393427
## x5  452.223413
```

```r
test_run4$b
```

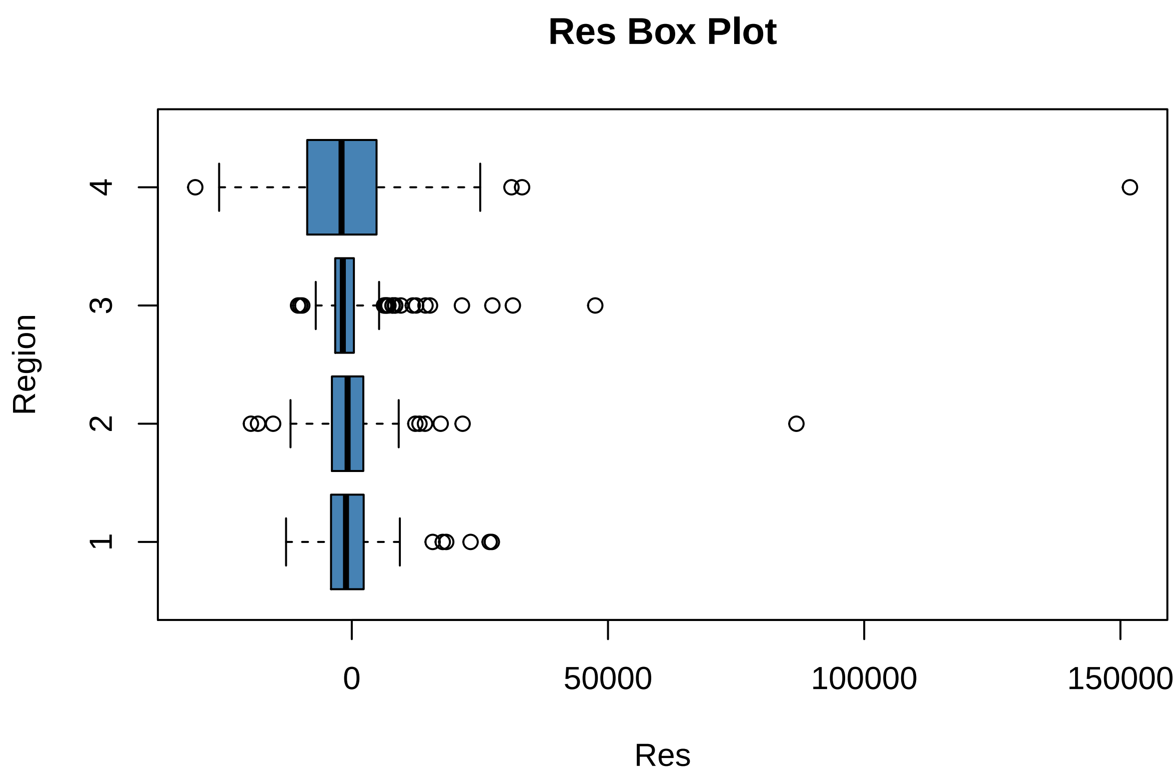
```
##           [,1]
##    138568.3499
## x1   1007.2152
## x2  -1809.1304
## x3    798.7147
## x4   -885.9512
## x5  -2494.6207
```

```r
e_generator = function(dat){
  outlist = lapply(1:4, function(i){
    loader =regional_data_generator2(dat= dat,region =i)
    y = loader$y
    x = loader$x
    n = nrow(y)
    I = diag(rep(1,n))
    H = x %*% solve(t(x)%*%x) %*% t(x)
    e = (I-H) %*% y
  })
  names(outlist)= c('Region1_error','Region2_error','Region3_error','Region4_error')

  r1 = cbind(outlist$Region1_error,rep(1,length(outlist$Region1_error)))
  r2 = cbind(outlist$Region2_error,rep(2,length(outlist$Region2_error)))
  r3 = cbind(outlist$Region3_error,rep(3,length(outlist$Region3_error)))
  r4 = cbind(outlist$Region4_error,rep(4,length(outlist$Region4_error)))
  r_dat = as.data.frame(rbind(r1,r2,r3,r4))
  colnames(r_dat) = c('Res','Region')

  return(r_dat)
}
```

```
r_dat = e_generator(dat=dat)
par(mfrow=c(1,1))
boxplot(r_dat$Res ~ r_dat$Region,
        col='steelblue',
        main='Res Box Plot',
        xlab='Res',
        ylab='Region',
        horizontal=T)
```



All four models have normal(symmetric) distribution. Model 1, 2, and 3 each have an outlier. Model 1 has the most longest interquartile range which means the data is most dispersed. The median of model 4 is within model 1, showing that these two models do not have a difference. The median of model 2 is within model 3, so these two models are similar.

Q2)

```
dat = read.delim2('Demographic.txt',header=F,sep ='')
names(dat) = c('id','county','state','land_area','total_population','percent_population_18to34','percen
               'No_physicians','No_hospital_beds','crimes','highschool','bachelor',
               'below_poverty','unemployment','per_cap_income','total_income','geo_region')

dat1 = dat[,c('crimes','bachelor','highschool','percent_population_18to34','percent_population_65orOlde
names(dat1)=c('Y','X1','X2','X3','X4','X5','region')
for (i in 1:6){
  dat1[,i] = as.numeric(dat1[,i])
}
region1=dat1[dat1$region==1, c('Y','X1','X2','X3','X4','X5')]
```

```r
region2=dat1[dat1$region==2, c('Y','X1','X2','X3','X4','X5')]
region3=dat1[dat1$region==3, c('Y','X1','X2','X3','X4','X5')]
region4=dat1[dat1$region==4, c('Y','X1','X2','X3','X4','X5')]

attach(dat1)

library(leaps)

nn<-length(region1[,1])
vars <- c("Y","X1","X2","X3","X4","X5")
N <- list(0,1,2,3,4,5)
COMB <- sapply(N, function(m) combn(x=vars[2:6], m))
COMB2 <- list()
k=0
exlmf<-lm(Y~X1+X2+X3+X4+X5)
SSEF<-deviance(aov(exlmf))
MSEF<-deviance(aov(exlmf))/(n-5)
```

```
## Error in eval(expr, envir, enclos): object 'n' not found
```

```r
res.table<-NULL
res.table0<-NULL
for(i in seq(COMB))
{
  tmp <- COMB[[i]]
  for(j in seq(ncol(tmp)))
  {k <- k + 1
  if(length(tmp)==0) COMB2[[k]] <- formula(paste("Y~1"))
  if(length(tmp)>0) COMB2[[k]] <- formula(paste("Y", "~",
                                             paste(tmp[,j], collapse=" + ")))
  exlm0<-lm(COMB2[[k]],data=dat1)
  exlm<-summary(exlm0)
  np<-length(tmp[,1])+1
  R2p<-exlm$r.squared
  R2ap<-exlm$adj.r.squared
  SSEp<-deviance(aov(exlm0))
  Cp<-SSEp/MSEF-(nn-2*(np))
  AICp<-nn*log(SSEp)-nn*log(nn)+2*np
  BICp<-nn*log(SSEp)-nn*log(nn)+log(nn)*np
  PRESSp<-sum((resid(exlm0)/(1 - lm.influence(exlm0)$hat))^2)
  res.table<-rbind(res.table,c(noquote(paste(tmp[,j], collapse=" + ")),
                              format(round(c(np,R2p,R2ap,Cp,AICp,BICp,PRESSp),3), nsmall = 3)))
  res.table0<-rbind(res.table0,c(np,R2p,R2ap,Cp,AICp,BICp,PRESSp))
  #names(res.table0) = c('Model','np','R2p','R2ap','Cp','AICp','BICp','PRESSp')
  }
}
```

```
## Error in eval(expr, envir, enclos): object 'MSEF' not found
```

```r
library(xtable)
tt = as.data.frame(res.table)
names(tt) = c('Model','np','R2p','R2ap','Cp','AICp','BICp','PRESSp')
```

```
## Error in names(tt) = c("Model", "np", "R2p", "R2ap", "Cp", "AICp", "BICp", : 'names' attribute [8] mu
```

```r
nn<-length(region2[,1])
vars <- c("Y","X1","X2","X3","X4","X5")
N <- list(0,1,2,3,4,5)
COMB <- sapply(N, function(m) combn(x=vars[2:6], m))
COMB2 <- list()
k=0
exlmf<-lm(Y~X1+X2+X3+X4+X5)
SSEF<-deviance(aov(exlmf))
MSEF<-deviance(aov(exlmf))/(n-5)
```

```
## Error in eval(expr, envir, enclos): object 'n' not found
```

```r
res.table<-NULL
res.table0<-NULL
for(i in seq(COMB))
{
  tmp <- COMB[[i]]
  for(j in seq(ncol(tmp)))
  {k <- k + 1
  if(length(tmp)==0) COMB2[[k]] <- formula(paste("Y~1"))
  if(length(tmp)>0) COMB2[[k]] <- formula(paste("Y", "~",
                                               paste(tmp[,j], collapse=" + ")))
  exlm0<-lm(COMB2[[k]],data=dat1)
  exlm<-summary(exlm0)
  np<-length(tmp[,1])+1
  R2p<-exlm$r.squared
  R2ap<-exlm$adj.r.squared
  SSEp<-deviance(aov(exlm0))
  Cp<-SSEp/MSEF-(nn-2*(np))
  AICp<-nn*log(SSEp)-nn*log(nn)+2*np
  BICp<-nn*log(SSEp)-nn*log(nn)+log(nn)*np
  PRESSp<-sum((resid(exlm0)/(1 - lm.influence(exlm0)$hat))^2)
  res.table<-rbind(res.table,c(noquote(paste(tmp[,j], collapse=" + ")),
                               format(round(c(np,R2p,R2ap,Cp,AICp,BICp,PRESSp),3), nsmall = 3)))
  res.table0<-rbind(res.table0,c(np,R2p,R2ap,Cp,AICp,BICp,PRESSp))
  #names(res.table0) = c('Model','np','R2p','R2ap','Cp','AICp','BICp','PRESSp')
  }
}
```

```
## Error in eval(expr, envir, enclos): object 'MSEF' not found
```

```r
library(xtable)
tt_2 = as.data.frame(res.table)
names(tt_2) = c('Model','np','R2p','R2ap','Cp','AICp','BICp','PRESSp')
```

```
## Error in names(tt_2) = c("Model", "np", "R2p", "R2ap", "Cp", "AICp", "BICp", : 'names' attribute [8]
```

```r
nn<-length(region3[,1])
vars <- c("Y","X1","X2","X3","X4","X5")
N <- list(0,1,2,3,4,5)
```

```
COMB <- sapply(N, function(m) combn(x=vars[2:6], m))
COMB2 <- list()
k=0
exlmf<-lm(Y~X1+X2+X3+X4+X5)
SSEF<-deviance(aov(exlmf))
MSEF<-deviance(aov(exlmf))/(n-5)
```

```
## Error in eval(expr, envir, enclos): object 'n' not found
```

```
res.table<-NULL
res.table0<-NULL
for(i in seq(COMB))
{
  tmp <- COMB[[i]]
  for(j in seq(ncol(tmp)))
  {k <- k + 1
  if(length(tmp)==0) COMB2[[k]] <- formula(paste("Y~1"))
  if(length(tmp)>0) COMB2[[k]] <- formula(paste("Y", "~",
                                             paste(tmp[,j], collapse=" + ")))
  exlm0<-lm(COMB2[[k]],data=dat1)
  exlm<-summary(exlm0)
  np<-length(tmp[,1])+1
  R2p<-exlm$r.squared
  R2ap<-exlm$adj.r.squared
  SSEp<-deviance(aov(exlm0))
  Cp<-SSEp/MSEF-(nn-2*(np))
  AICp<-nn*log(SSEp)-nn*log(nn)+2*np
  BICp<-nn*log(SSEp)-nn*log(nn)+log(nn)*np
  PRESSp<-sum((resid(exlm0)/(1 - lm.influence(exlm0)$hat))^2)
  res.table<-rbind(res.table,c(noquote(paste(tmp[,j], collapse=" + ")),
                              format(round(c(np,R2p,R2ap,Cp,AICp,BICp,PRESSp),3), nsmall = 3)))
  res.table0<-rbind(res.table0,c(np,R2p,R2ap,Cp,AICp,BICp,PRESSp))
  #names(res.table0) = c('Model','np','R2p','R2ap','Cp','AICp','BICp','PRESSp')
  }
}
```

```
## Error in eval(expr, envir, enclos): object 'MSEF' not found
```

```
library(xtable)
tt_3 = as.data.frame(res.table)
names(tt_3) = c('Model','np','R2p','R2ap','Cp','AICp','BICp','PRESSp')
```

```
## Error in names(tt_3) = c("Model", "np", "R2p", "R2ap", "Cp", "AICp", "BICp", : 'names' attribute [8]
```

```
nn<-length(region4[,1])
vars <- c("Y","X1","X2","X3","X4","X5")
N <- list(0,1,2,3,4,5)
COMB <- sapply(N, function(m) combn(x=vars[2:6], m))
COMB2 <- list()
k=0
exlmf<-lm(Y~X1+X2+X3+X4+X5)
SSEF<-deviance(aov(exlmf))
MSEF<-deviance(aov(exlmf))/(n-5)
```

```
## Error in eval(expr, envir, enclos): object 'n' not found
```

```
res.table<-NULL
res.table0<-NULL
for(i in seq(COMB))
{
  tmp <- COMB[[i]]
  for(j in seq(ncol(tmp)))
  {k <- k + 1
  if(length(tmp)==0) COMB2[[k]] <- formula(paste("Y~1"))
  if(length(tmp)>0) COMB2[[k]] <- formula(paste("Y", "~",
                                              paste(tmp[,j], collapse=" + ")))
  exlm0<-lm(COMB2[[k]],data=dat1)
  exlm<-summary(exlm0)
  np<-length(tmp[,1])+1
  R2p<-exlm$r.squared
  R2ap<-exlm$adj.r.squared
  SSEp<-deviance(aov(exlm0))
  Cp<-SSEp/MSEF-(nn-2*(np))
  AICp<-nn*log(SSEp)-nn*log(nn)+2*np
  BICp<-nn*log(SSEp)-nn*log(nn)+log(nn)*np
  PRESSp<-sum((resid(exlm0)/(1 - lm.influence(exlm0)$hat))^2)
  res.table<-rbind(res.table,c(noquote(paste(tmp[,j], collapse=" + ")),
                              format(round(c(np,R2p,R2ap,Cp,AICp,BICp,PRESSp),3), nsmall = 3)))
  res.table0<-rbind(res.table0,c(np,R2p,R2ap,Cp,AICp,BICp,PRESSp))
  #names(res.table0) = c('Model','np','R2p','R2ap','Cp','AICp','BICp','PRESSp')
  }
}
```

```
## Error in eval(expr, envir, enclos): object 'MSEF' not found
```

```
library(xtable)
tt_4 = as.data.frame(res.table)
names(tt_4) = c('Model','np','R2p','R2ap','Cp','AICp','BICp','PRESSp')
```

```
## Error in names(tt_4) = c("Model", "np", "R2p", "R2ap", "Cp", "AICp", "BICp", : 'names' attribute [8]
```

Region 1: X1 + X2 has the lowest AIC value. B0 has the lowest SBC value. Region 2: X1 + X2 has the lowest AIC value. B0 has the lowest SBC value. Region 3: X1 + X2 has the lowest AIC value. B0 has the lowest SBC value. Region 4: X1 + X2 has the lowest AIC value. B0 has the lowest SBC value.

Q3) Region 1: X1 + X2 model: B0 has a confidence interval of (331669.753, 699115.906). B1 has a confidence interval of (2372.152, 6923.484) and B2 has a confidence interval of (-10478.779, -4821.719). B0 model: (11947.87, 34224.19)

```
model1<-lm(region1$Y~region1$X1+region1$X2)
confint(model1, level=0.9)
```

```
##                     5 %       95 %
## (Intercept) 331669.753 699115.906
## region1$X1    2372.152   6923.484
## region1$X2  -10478.779  -4821.719
```

```r
summary(model1)
```

```
##
## Call:
## lm(formula = region1$Y ~ region1$X1 + region1$X2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -75993 -19688  -3162  10859 575740
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   515393     110661   4.657 9.87e-06 ***
## region1$X1      4648       1371   3.391 0.000999 ***
## region1$X2     -7650       1704  -4.490 1.91e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62730 on 100 degrees of freedom
## Multiple R-squared:  0.168,  Adjusted R-squared:  0.1514
## F-statistic:  10.1 on 2 and 100 DF,  p-value: 0.0001013
```

```r
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: region1$Y
##             Df     Sum Sq    Mean Sq F value    Pr(>F)
## region1$X1   1 1.2330e+08 1.2330e+08  0.0313    0.8599
## region1$X2   1 7.9354e+10 7.9354e+10 20.1636 1.909e-05 ***
## Residuals  100 3.9355e+11 3.9355e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
model2<-lm(region1$Y~1)
confint(model2, level=0.9)
```

```
##                  5 %     95 %
## (Intercept) 11947.87 34224.19
```

```r
summary(model2)
```

```
##
## Call:
## lm(formula = region1$Y ~ 1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -21689 -18926 -13999   -311 657880
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    23086       6710   3.441 0.000843 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68100 on 102 degrees of freedom
```

```r
anova(model2)
```

```
## Analysis of Variance Table
##
## Response: region1$Y
##            Df     Sum Sq    Mean Sq F value Pr(>F)
## Residuals 102 4.7303e+11 4637509423
```

X1 + X2 model: Decision rule: - If abs(t_) <= t(l - 0.01/2; n - p), conclude H0 Conclusion: - p-value: 0.000999 and 1.91e-05 - Based on our decision rule, we reject H0. Decision rule: - Reject H0 if p-value < 0.05 or - Reject H0 if F_ > F(l - 0.05; p - 1, n - p) Conclusion: F* is 10.1 and p-value is 0.0001013. Therefore, we reject H0 based on our decision rule.

B0 model: No need to do a test since this means the null hypothesis is true (no parameters).

Region 2: X1 + X2 model: B0 has a confidence interval of (278838.013, 574755.685). B1 has a confidence interval of (2667.359, 5893.015) and B2 has a confidence interval of (-8344.590, -3992.670). B0 model: (14220.76, 29340.73)

```r
model3<-lm(region2$Y~region2$X1+region2$X2)
confint(model3, level=0.9)
```

```
##                     5 %       95 %
## (Intercept) 278838.013 574755.685
## region2$X1    2667.359   5893.015
## region2$X2   -8344.590  -3992.670
```

```r
summary(model3)
```

```
##
## Call:
## lm(formula = region2$Y ~ region2$X1 + region2$X2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -59513 -16980  -4658   8583 365328
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 426796.8    89158.9   4.787 5.57e-06 ***
## region2$X1    4280.2      971.9   4.404 2.57e-05 ***
## region2$X2   -6168.6     1311.2  -4.705 7.79e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43280 on 105 degrees of freedom
## Multiple R-squared:  0.1802, Adjusted R-squared:  0.1645
## F-statistic: 11.54 on 2 and 105 DF,  p-value: 2.956e-05
```

```r
anova(model3)
```

```
## Analysis of Variance Table
##
## Response: region2$Y
##              Df     Sum Sq    Mean Sq F value    Pr(>F)
## region2$X1    1 1.7638e+09 1.7638e+09  0.9416    0.3341
## region2$X2    1 4.1458e+10 4.1458e+10 22.1324 7.788e-06 ***
## Residuals   105 1.9668e+11 1.8732e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
model4<-lm(region2$Y~1)
confint(model4, level=0.9)
```

```
##                  5 %     95 %
## (Intercept) 14220.76 29340.73
```

```r
summary(model4)
```

```
##
## Call:
## lm(formula = region2$Y ~ 1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -21218 -16586 -11944  -2337 415155
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    21781       4556    4.78 5.61e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 47350 on 107 degrees of freedom
```

```r
anova(model4)
```

```
## Analysis of Variance Table
##
## Response: region2$Y
##            Df     Sum Sq    Mean Sq F value Pr(>F)
## Residuals 107 2.3991e+11 2242116022
```

X1 + X2 model: Decision rule: - If abs(t*) <= t(l - 0.01/2; n - p), conclude H0 Conclusion: - p-value: 2.57e-05 and 7.79e-06 - Based on our decision rule, we reject H0. Decision rule: - Reject H0 if p-value < 0.05 or - Reject H0 if F* > F(l - 0.05; p - 1, n - p) Conclusion: F* is 11.54 and p-value is 2.956e-05. Therefore, we reject H0 based on our decision rule.

B0 model: No need to do a test since this means the null hypothesis is true (no parameters).

Region 3: X1 + X2 model: B0 has a confidence interval of (13685.6700, 135109.52034). B1 has a confidence interval of (462.7984, 2291.32024) and B2 has a confidence interval of (-1988.4944, -44.80433). B0 model: (21832.74, 32149.2)

```r
model5<-lm(region3$Y~region3$X1+region3$X2)
confint(model5, level=0.9)
```

```
##                   5 %         95 %
## (Intercept) 13685.6700 135109.52034
## region3$X1    462.7984   2291.32024
## region3$X2  -1988.4944    -44.80433
```

```r
summary(model5)
```

```
##
## Call:
## lm(formula = region3$Y ~ region3$X1 + region3$X2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -44935 -17721 -11078   1027 221675
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  74397.6    36680.7   2.028   0.0443 *
## region3$X1    1377.1      552.4   2.493   0.0138 *
## region3$X2   -1016.6      587.2  -1.731   0.0854 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37900 on 149 degrees of freedom
## Multiple R-squared:  0.04012,    Adjusted R-squared:  0.02723
## F-statistic: 3.114 on 2 and 149 DF,  p-value: 0.04734
```

```r
anova(model5)
```

```
## Analysis of Variance Table
##
## Response: region3$Y
##             Df     Sum Sq    Mean Sq F value  Pr(>F)
## region3$X1   1 4.6386e+09 4638631266  3.2295 0.07435 .
## region3$X2   1 4.3060e+09 4305996828  2.9979 0.08544 .
## Residuals  149 2.1401e+11 1436327617
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
model6<-lm(region3$Y~1)
confint(model6, level=0.9)
```

```
##                  5 %    95 %
## (Intercept) 21832.74 32149.2
```

```r
summary(model6)
```

```
##
## Call:
## lm(formula = region3$Y ~ 1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -24051 -19471 -12546    743 226535
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    26991       3117    8.66 6.67e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38430 on 151 degrees of freedom
```

```r
anova(model6)
```

```
## Analysis of Variance Table
##
## Response: region3$Y
##            Df     Sum Sq    Mean Sq F value Pr(>F)
## Residuals 151 2.2296e+11 1476539357
```

X1 + X2 model: Decision rule: - If abs(t*) <= t(l - 0.01/2; n - p), conclude H0 Conclusion: - p-value: 0.0138 and 0.0854 - Based on our decision rule, we reject H0 for X1, but we conclude H0 for X2. Decision rule: - Reject H0 if p-value < 0.05 or - Reject H0 if F *>* F(l - 0.05; p - 1, n - p) Conclusion: F* is 3.114 and p-value is 0.04734. Therefore, we reject H0 based on our decision rule.

B0 model: No need to do a test since this means the null hypothesis is true (no parameters).

Region 4: X1 + X2 model: B0 has a confidence interval of (60623.396, 433810.9147). B1 has a confidence interval of (810.656, 7058.3043) and B2 has a confidence interval of (-6563.551, -808.7508). B0 model: (24289.37, 56134.08)

```r
model7<-lm(region4$Y~region4$X1+region4$X2)
confint(model7, level=0.9)
```

```
##                   5 %        95 %
## (Intercept) 60623.396 433810.9147
## region4$X1    810.656   7058.3043
## region4$X2  -6563.551   -808.7508
```

```r
summary(model7)
```

```
##
## Call:
## lm(formula = region4$Y ~ region4$X1 + region4$X2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -81236 -28955 -14033   3792 612011
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   247217     112021   2.207   0.0304 *
## region4$X1      3934       1875   2.098   0.0393 *
## region4$X2     -3686       1727  -2.134   0.0362 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 82190 on 74 degrees of freedom
## Multiple R-squared:  0.06575,    Adjusted R-squared:  0.0405
## F-statistic: 2.604 on 2 and 74 DF,  p-value: 0.08076
```

```
anova(model7)
```

```
## Analysis of Variance Table
##
## Response: region4$Y
##            Df     Sum Sq    Mean Sq F value  Pr(>F)
## region4$X1  1 4.4190e+09 4.4190e+09  0.6542 0.42122
## region4$X2  1 3.0760e+10 3.0760e+10  4.5535 0.03617 *
## Residuals  74 4.9989e+11 6.7553e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model8<-lm(region4$Y~1)
confint(model8, level=0.9)
```

```
##                  5 %     95 %
## (Intercept) 24289.37 56134.08
```

```
summary(model8)
```

```
##
## Call:
## lm(formula = region4$Y ~ 1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -36333 -31625 -24121   1068 648724
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    40212       9562   4.205 7.05e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 83910 on 76 degrees of freedom
```

```
anova(model8)
```

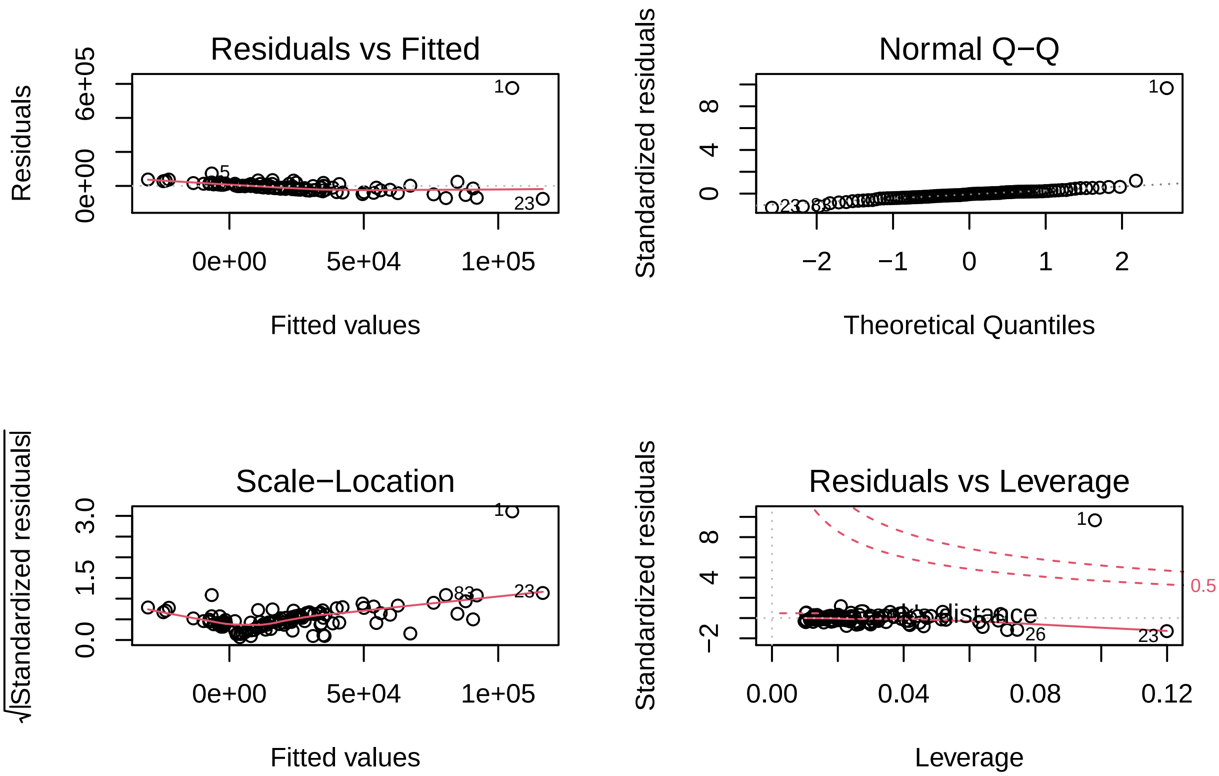
```
## Analysis of Variance Table
```

```
##
## Response: region4$Y
##            Df     Sum Sq    Mean Sq F value Pr(>F)
## Residuals 76 5.3507e+11 7040407811
```

X1 + X2 model: Decision rule: - If abs(t*) <= t(l - 0.01/2; n - p), conclude H0 Conclusion: - p-value: 0.0393 and 0.0362 - Based on our decision rule, we reject H0. Decision rule: - Reject H0 if p-value < 0.05 or - Reject H0 if F >* F(l - 0.05; p - 1, n - p) Conclusion: F* is 2.604 and p-value is 0.08076. Therefore, we reject H0 based on our decision rule.

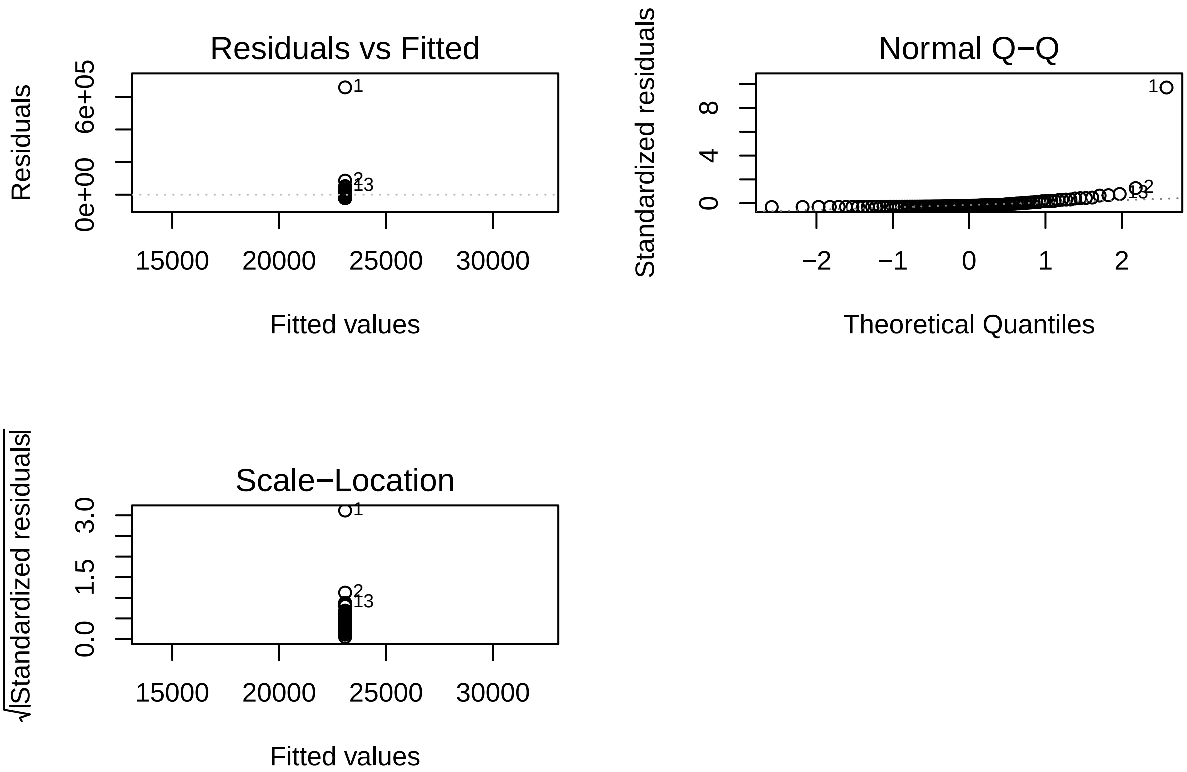
B0 model: No need to do a test since this means the null hypothesis is true (no parameters).

```r
par(mfrow = c(2, 2))
plot(model1)
```

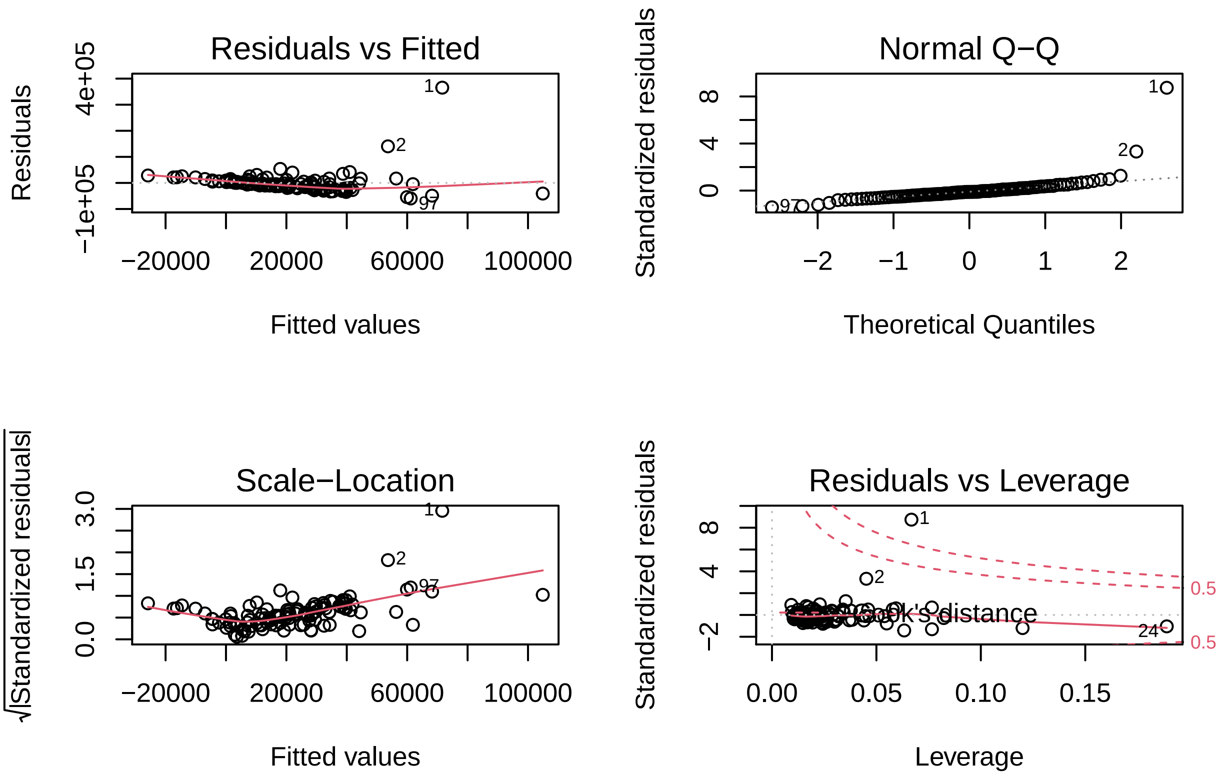



```r
par(mfrow = c(2, 2))
plot(model2)
```

```
## hat values (leverages) are all = 0.009708738
##  and there are no factor predictors; no plot no. 5
```

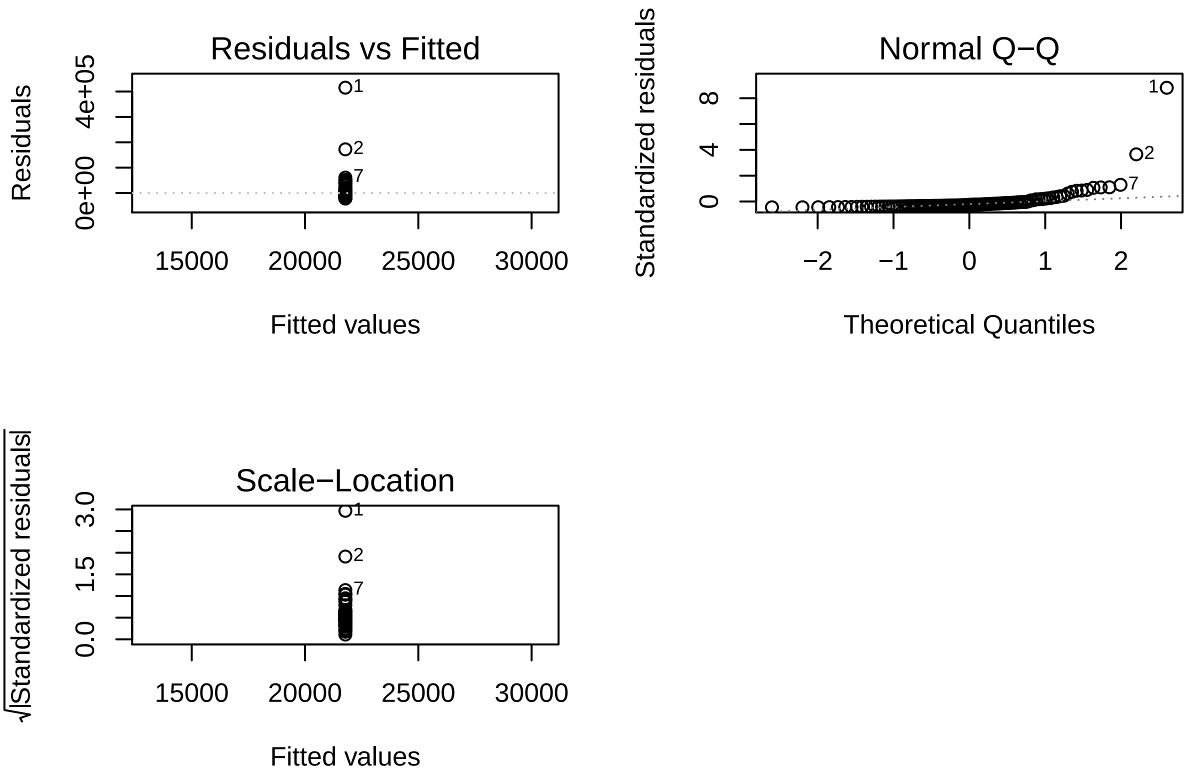



```
par(mfrow = c(2, 2))
plot(model3)
```

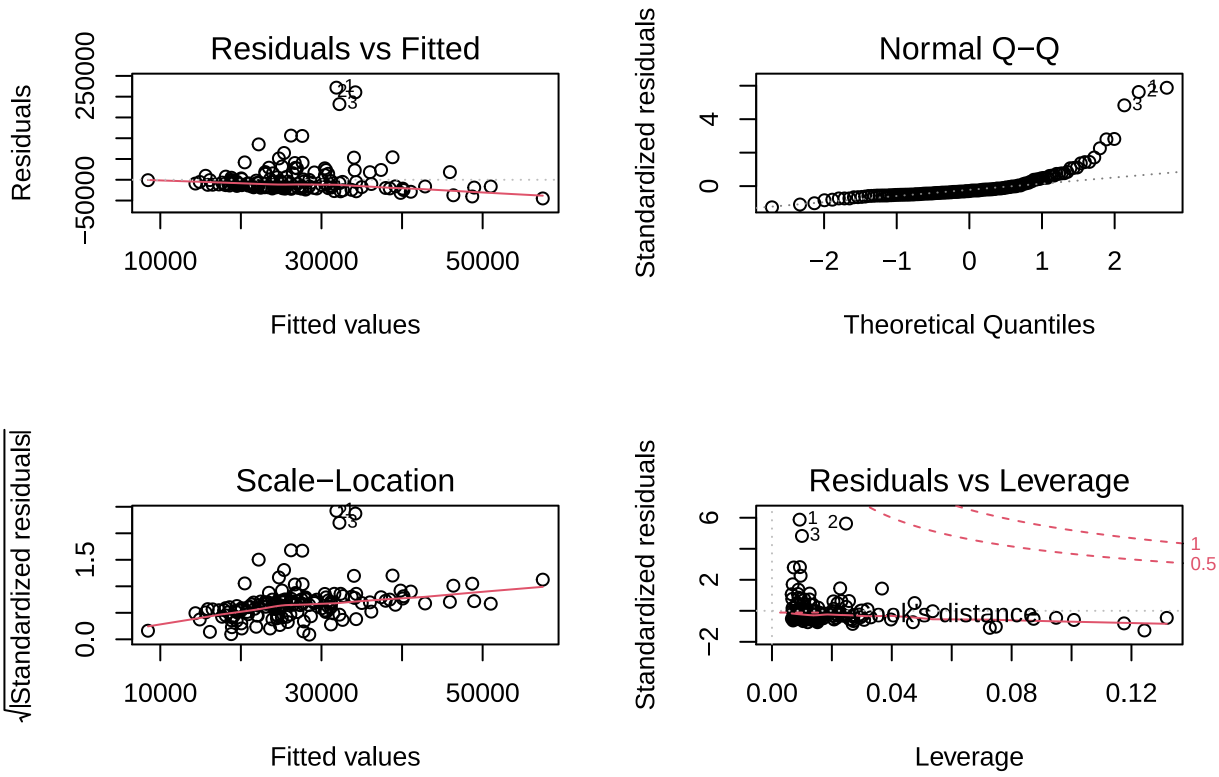



```r
par(mfrow = c(2, 2))
plot(model4)
```

```
## hat values (leverages) are all = 0.009259259
##  and there are no factor predictors; no plot no. 5
```

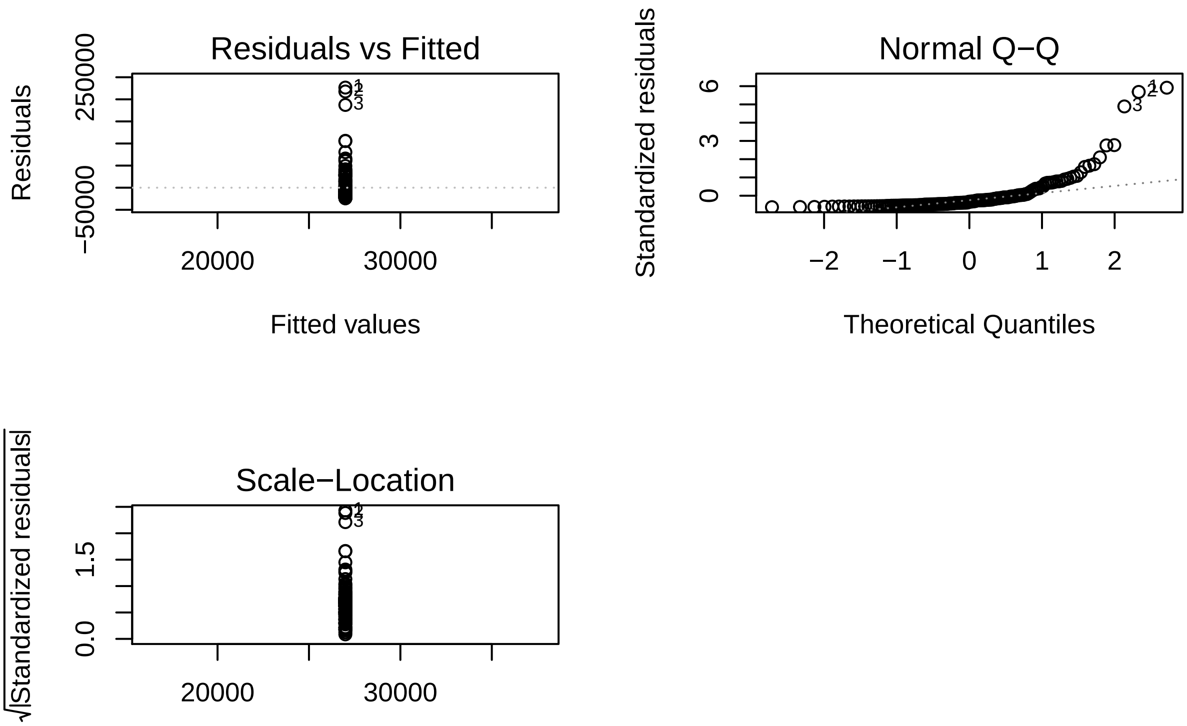



```
par(mfrow = c(2, 2))
plot(model5)
```

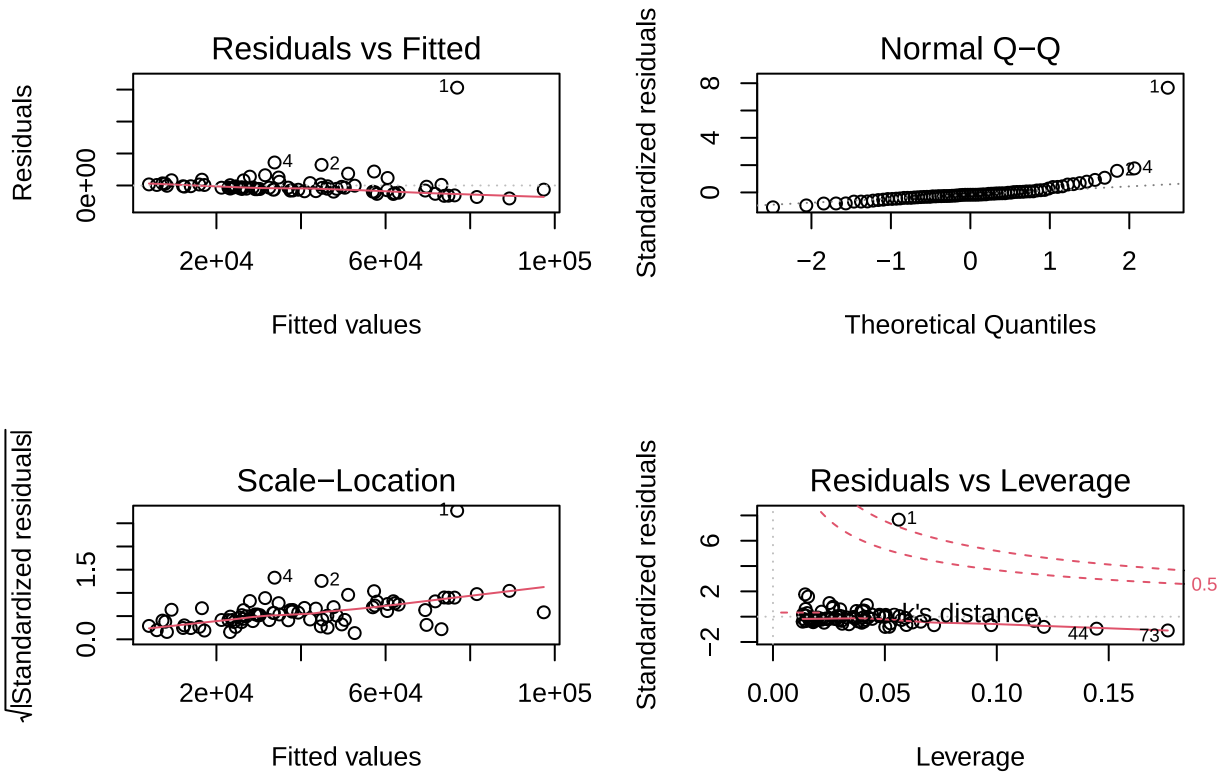



```r
par(mfrow = c(2, 2))
plot(model6)
```

```
## hat values (leverages) are all = 0.006578947
##  and there are no factor predictors; no plot no. 5
```

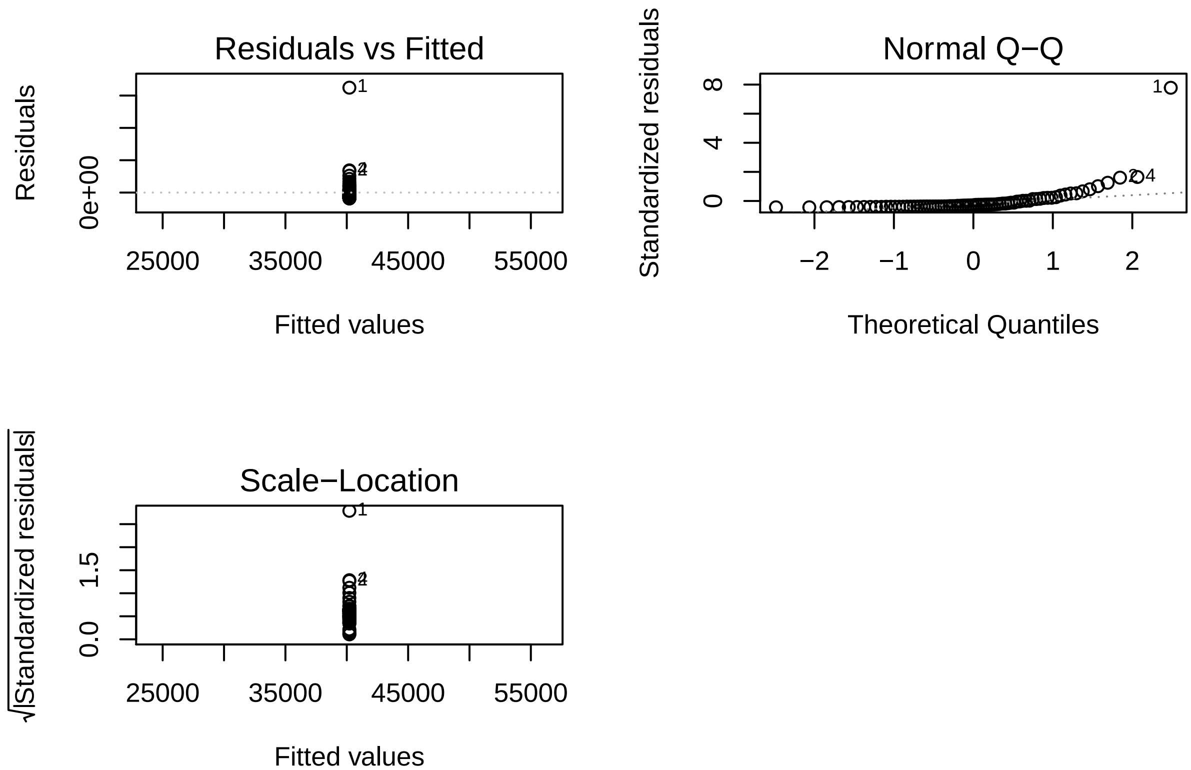



```r
par(mfrow = c(2, 2))
plot(model7)
```




```r
par(mfrow = c(2, 2))
plot(model8)
```

```
## hat values (leverages) are all = 0.01298701
##  and there are no factor predictors; no plot no. 5
```



X1+X2 models for all regions: follows the red line in the residual line. This suggests non-linear relationship in the data. Model 5 has the least pattern, suggesting model 5 is the cloest to a linear relationship. Model 1 has the best normality and model 5 deviates most from normality. They all have a reasonable heteroscedasticity. Model 1 has outliers of #1, #5, #23, #26, #83 (five outliers). Model 3 has outliers of #1, #2, #97 (three outliers). Model 5 has outliers of #1, #2, #3 (three outliers). Model 7 has outliers of #1, #2, #4, #44, #73 (five outliers). B0 model: We can clearly see that this model is an intercept-only model from the plots. It shows the same pattern as X1+X2 model. Model 6 deviates most from normality and model 2 shows the most normality. All models have three outliers.