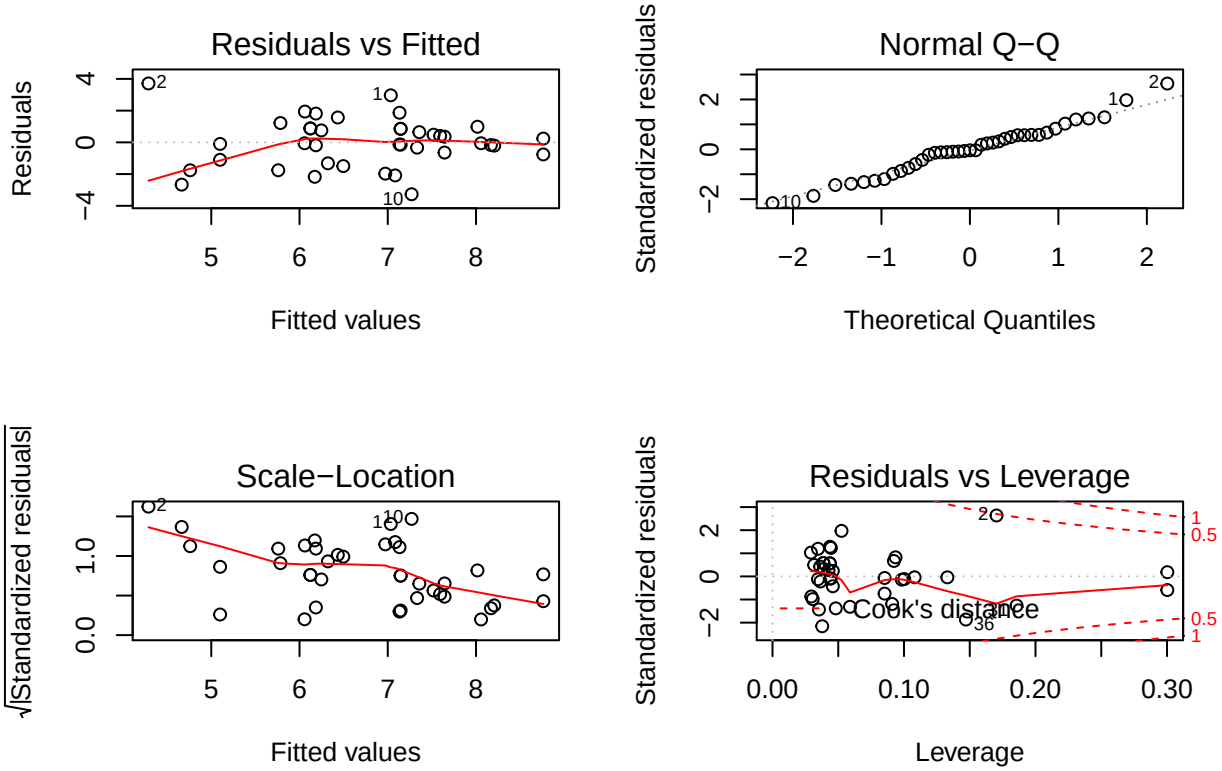
STAT425_HW4 Jinran Yang

1.

```
library(faraway)
data(happy)
lm<-lm(happy~money+work,data = happy)
par(mfrow = c(2,2))
plot(lm)
```



(1).As we can see from the Scale-Location plot above, the line tends to be a curve instead of a flat line, so constant variance assumption for the errors is false.

(2).As we can see from the Normal Q-Q plot above, the residuals are not a straight line so that the normality assumption is not well satisfied.

(3).

```
which(hatvalues(lm)>(2*3/39))#2p/n
```

```
##  2  6  7 31
##  2  6  7 31
```

Therefore, #2,6,7 and 31 observations can be considered as the large leverage points.

(4).

```
critval <- qt(0.05/(2*nobs(lm)), df=df.residual(lm)-1, lower=FALSE)#alpha/n
 which(abs(rstudent(lm)) > critval)
```

```
## named integer(0)
```

There is no outlier.

(5)

```
which(cooks.distance(lm)>=1)
```

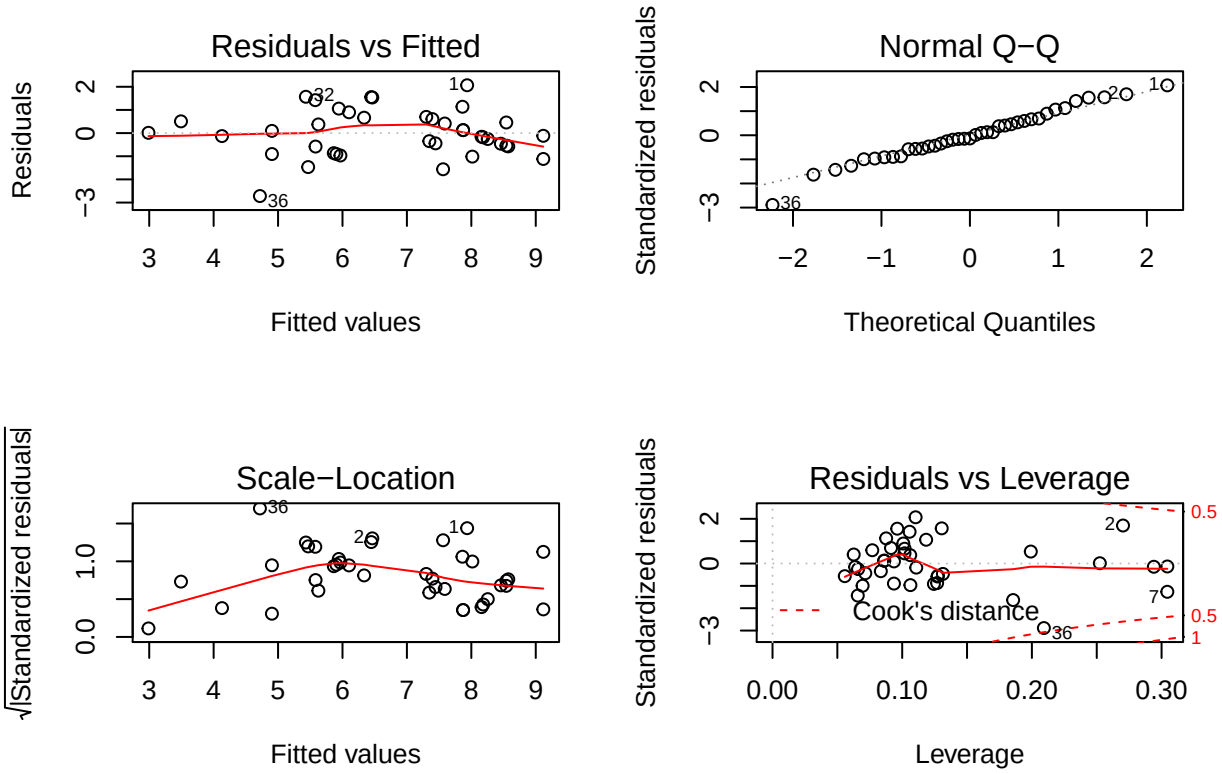
```
## named integer(0)
```

There is no influential point. And we can also get the same result from the `Residuals and Leverage` plot, because there is no observation outside the contour line of cook's distance equal to 1.

(6) We can get the structure of the relationship using the `Residual versus Fitted` plot. As it seems that the line is not very flat, the relationship between response and the predictors may not be simply linear.

New model

```
lm2<-lm(happy~.,data = happy)
par(mfrow = c(2,2))
plot(lm2)
```



(1)As we can see from the Scale-Location plot above, the line become much flatter than before which means the constant variance assumption is more satisfied in this model.

(2)As we can see from the Normal Q-Q plot above, the residuals looks like a straight line so that the normality assumption is satisfied.

(3)

```
which(hatvalues(lm2)>(2*5/39))
```

```
##  2  6  7 10
##  2  6  7 10
```

Therefore, #2,6,7 and #10 observations can be considered as the large leverage points.

(4)

```
critval <- qt(0.05/(2*nobs(lm2)), df=df.residual(lm2)-1, lower=FALSE)#alpha/n
 which(abs(rstudent(lm2)) >critval)
```

```
## named integer(0)
```

There is no outlier.

(5)

```
which(cooks.distance(lm)>=1)
```

```
## named integer(0)
```

There is no influential point. And we can also get the same result from the `Residuals and Leverage` graph, there is no observation outside the contour line of cook's distance is 1.

(6) We can get the structure of the relationship using the `Residual versus Fitted` plot. As it seems that the line is getting more flat, so that we can say that after adding more predictors in the model, the relationship between response and the predictors become more linear.

2.

(a)

```
data("seatpos")
full_model<-lm(hipcenter~.,data=seatpos)
summary(full_model)
```

```
##
## Call:
## lm(formula = hipcenter ~ ., data = seatpos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -73.827 -22.833  -3.678  25.017  62.337
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 436.43213  166.57162   2.620   0.0138 *
## Age           0.77572    0.57033   1.360   0.1843
## Weight        0.02631    0.33097   0.080   0.9372
## HtShoes      -2.69241    9.75304  -0.276   0.7845
## Ht            0.60134   10.12987   0.059   0.9531
## Seated        0.53375    3.76189   0.142   0.8882
## Arm          -1.32807    3.90020  -0.341   0.7359
## Thigh        -1.14312    2.66002  -0.430   0.6706
## Leg          -6.43905    4.71386  -1.366   0.1824
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.72 on 29 degrees of freedom
## Multiple R-squared:  0.6866, Adjusted R-squared:  0.6001
## F-statistic:  7.94 on 8 and 29 DF,  p-value: 1.306e-05
```

(b) According to the result above, no variable appear to be significant based on the individual t-tests, since all the p value of their t test are much larger than 0.05; but the overall F-test is significant which means

the overall model is significant.

(c)

```
vif(full_model)
```

```
##        Age     Weight    HtShoes         Ht     Seated        Arm
##   1.997931   3.647030 307.429378 333.137832   8.951054   4.496368
##      Thigh        Leg
##   2.762886   6.694291
```

As we can see from the above, `HtShoes` and `Ht` both have a very high VIF(much larger than 10) so that they might have a problem of collinearity.

(d)

```
reduced_model<-lm(hipcenter~ Age+Weight+Seated+Arm+Thigh+Leg,data=seatpos)
summary(reduced_model)
```

```
##
## Call:
## lm(formula = hipcenter ~ Age + Weight + Seated + Arm + Thigh +
##     Leg, data = seatpos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.296 -23.340  -5.672  24.183  74.065
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 409.00851  159.49517   2.564   0.0154 *
## Age           0.83110    0.52771   1.575   0.1254
## Weight       -0.03251    0.31254  -0.104   0.9178
## Seated       -1.73576    2.48225  -0.699   0.4896
## Arm          -2.00541    3.69731  -0.542   0.5914
## Thigh        -1.91970    2.24858  -0.854   0.3998
## Leg          -8.40876    3.91939  -2.145   0.0399 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.91 on 31 degrees of freedom
## Multiple R-squared:  0.6791, Adjusted R-squared:  0.617
## F-statistic: 10.94 on 6 and 31 DF,  p-value: 1.571e-06
```

(e) According to the result above, `Leg` is significant based on the individual t-tests for its coefficients. And overall F-test (for all of the variables together) is significant, since the p-value is much smaller than 0.05.

(f)

```
vif(reduced_model)
```

```
##      Age   Weight   Seated      Arm    Thigh      Leg
## 1.786192 3.396124 4.069626 4.219519 2.061632 4.832701
```

Using the threshold of 10, no variables in the new model have a VIF larger than 10.

4

(a)

```
library(MASS)
set.seed(1)
I<-diag(5)
J<-matrix(1,5,5)
Sigma<-0.8*I+0.2*J
Sigma
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]  1.0  0.2  0.2  0.2  0.2
## [2,]  0.2  1.0  0.2  0.2  0.2
## [3,]  0.2  0.2  1.0  0.2  0.2
## [4,]  0.2  0.2  0.2  1.0  0.2
## [5,]  0.2  0.2  0.2  0.2  1.0
```

```
X<-mvrnorm(n=100,rep(0,5),Sigma)
X<-cbind(rep(1,100),X)
head(X)
```

```
##      [,1]       [,2]       [,3]       [,4]       [,5]       [,6]
## [1,]    1 -0.1204211 -0.5633121  1.1065356  0.9448524  0.5117065
## [2,]    1 -0.0764933 -1.2532929  1.1986881 -1.4335690  1.0137371
## [3,]    1 -0.2273602 -0.6379227 -0.2962704  2.0062845  1.6621546
## [4,]    1 -0.8307455 -0.4944209 -1.1735926 -1.1671208 -1.1199627
## [5,]    1 -0.7213724  0.4015845 -0.2829439  1.5688489 -1.9546404
## [6,]    1  1.9061108 -1.4864476 -0.7389317  0.9496307  1.8310430
```

(b)

```
set.seed(1)
error_normal<-rnorm(90,mean = 0,sd=1)
error_t<-rt(10, 100-6)
error<-c(error_normal,error_t)
index<-sample(1:100,size = 100)#random shuffle
for (i in 1:100){
  error[i]<-error[index[i]]
}
error<-as.vector(error)
beta<-matrix(c(0, 1, -1, 1, -1, 0),6,1)
Y<-X%*%beta+error
```

(c)

```
X_model<-X[,-1]
mlm<-lm(Y~X_model)
summary(mlm)
```

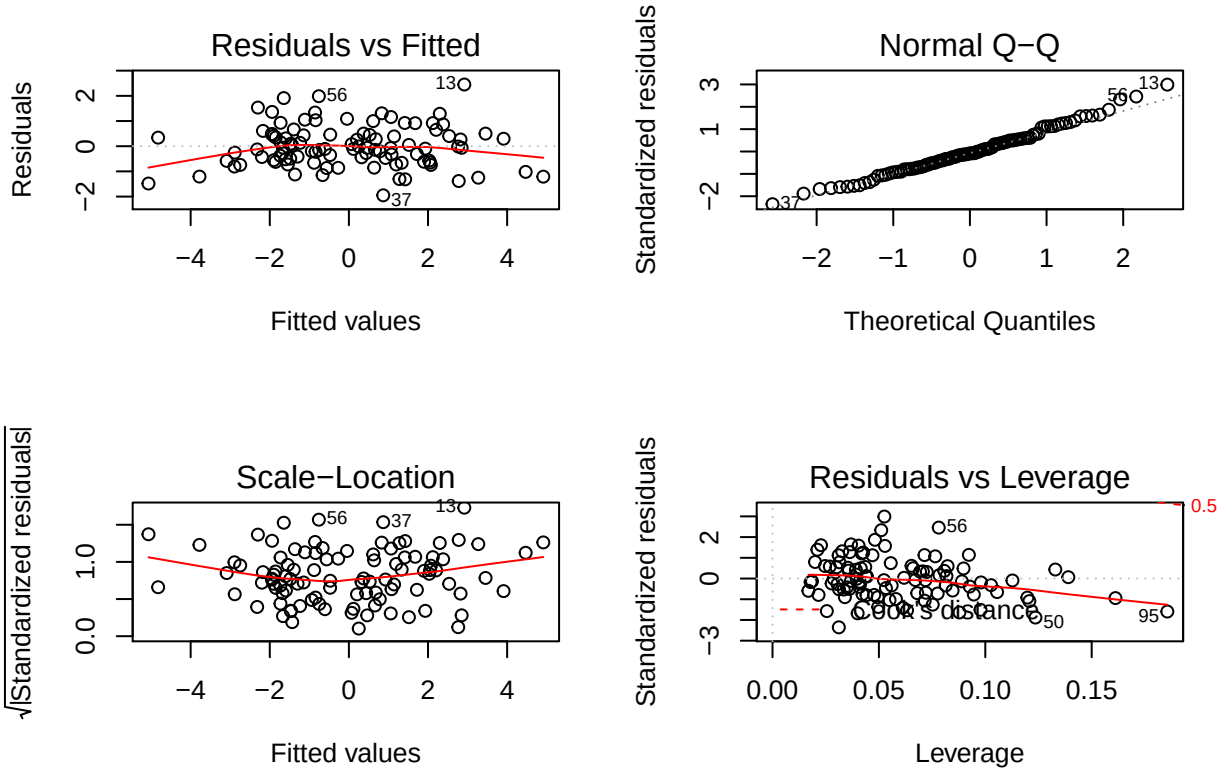
```
##
## Call:
## lm(formula = Y ~ X_model)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9499 -0.5750 -0.0692  0.4640  2.4491
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.06325    0.08504   0.744    0.459
## X_model1     0.88211    0.09413   9.371 3.99e-15 ***
## X_model2    -0.92145    0.08145 -11.313  < 2e-16 ***
## X_model3     1.10335    0.08979  12.289  < 2e-16 ***
## X_model4    -0.95422    0.08959 -10.651  < 2e-16 ***
## X_model5     0.06816    0.08949   0.762    0.448
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8411 on 94 degrees of freedom
## Multiple R-squared:  0.8536, Adjusted R-squared:  0.8458
## F-statistic: 109.6 on 5 and 94 DF,  p-value: < 2.2e-16
```

According to the result above, X1, X2, X3 and X4 all appear to be significant based on the individual t-tests, since all the p value of their t test are much smaller than 0.05. X5 is the only one which is not significant, and this is because when we generate the y, the beta corresponds to X5 is 0; The overall F-test is significant which means the overall model is significant.

(d-1) Check the constant variance assumption for the errors.

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



As we can see from the Scale-Location plot above, the line is flat overall,it just sinks a little in the middle, so that the constant variance assumption is overall satisfied in this model.

(d-2) Check the normality assumption.
As we can see from the Normal Q-Q plot above, the residuals looks like a straight line so that the normality assumption is satisfied.

(d-3) Check for large leverage points

```
which(hatvalues(mlm)> 2*6/100)
```

```
##  6 32 46 50 61 95
##  6 32 46 50 61 95
```

Therefore, #6,32, 46, 50, 61, and #95 observations can be considered as the large leverage points.

(d-4) Check for outliers.

```
critval <- qt(0.05/(2*nobs(mlm)), df=df.residual(mlm)-1, lower=FALSE)#alpha/n
 which(abs(rstudent(mlm)) > critval)
```

```
## named integer(0)
```

Based on the result above, there is no outlier.

(d-5) Check for influential points.

```
which(cooks.distance(mlm)>=1)
```

```
## named integer(0)
```

Based on the result above, there is no influential point.

(d-6) Check the structure of the relationship between the predictors and the response

We can get the structure of the relationship using the `Residual versus Fitted` plot. As it seems that the line is not vary flat, so that the relationship between response and the predictors may not be simply linear.