Assignment 2 Question 2 & 3

2.

a).

```
# import data
x1 <-c(4.49,3.04,3.94,2.63,4.55,3.88,2.92,2.82,3.17,2.91)
x2 <-c(2.92,4.33,4.27,1.92,2.47,2.36,3.21,4.22,1.80,2.35)
y <-c(-5.32,-9.24,-5.89,1.15,-1.47,1.91,-3.99,-6.82,1.49,-0.89)

# combine X matrix
X <- cbind(rep(1,10),x1,x2)

# Compute the covariance for the least squares regression estimators
diag(solve(t(X) %*% X)*2)
```

```
##                  x1        x2
## 7.5752722 0.4348632 0.2317155
```

So the covariance for the least squares regression estimators are 7.5752722, 0.4348632, and 0.2317155.

b).

```r
# estimate the least squares regression line
reg <- lm(y ~ x1+x2)
reg_sum <- summary(reg)

# find the estimate of V(beta-hat)
reg_sum$coefficients[,2]^2
```

```
## (Intercept)          x1          x2
##  10.2386726   0.5877574   0.3131846
```

So the estimate of V(beta-hat) are 10.2386726, 0.5877574, and 0.3131846.

c).

```
# compute hat matrix
H <- X %*% solve(t(X) %*% X) %*% t(X)

# compute the covariance matrix of the residuals
(evar <- (diag(10)-H)*2)
```

```
##              [,1]          [,2]       [,3]          [,4]        [,5]
##  [1,]  1.31620840 -0.0111751586 -0.4239149  0.1626592689 -0.71390443
##  [2,] -0.01117516  1.3222765078 -0.5152385 -0.0005628451  0.13711355
##  [3,] -0.42391486 -0.5152384606  1.2951190  0.3076598794 -0.30178685
##  [4,]  0.16265927 -0.0005628451  0.3076599  1.2403645336  0.07000078
##  [5,] -0.71390443  0.1371135484 -0.3017868  0.0700007844  1.20796776
##  [6,] -0.40754320  0.0638204487 -0.1138718 -0.1987124255 -0.48224196
##  [7,]  0.03728899 -0.3517406584 -0.1490973 -0.3279756195  0.07203972
##  [8,]  0.08899089 -0.6790200204 -0.4312100 -0.1075498945  0.23074070
##  [9,] -0.08547365  0.1228168179  0.2192572 -0.5960232666 -0.20254325
## [10,]  0.03686373 -0.0882901796  0.1130833 -0.5498604151 -0.01738603
##              [,6]        [,7]        [,8]        [,9]       [,10]
##  [1,] -0.40754320  0.03728899  0.08899089 -0.08547365  0.03686373
##  [2,]  0.06382045 -0.35174066 -0.67902002  0.12281682 -0.08829018
##  [3,] -0.11387180 -0.14909734 -0.43121001  0.21925718  0.11308330
##  [4,] -0.19871243 -0.32797562 -0.10754989 -0.59602327 -0.54986042
##  [5,] -0.48224196  0.07203972  0.23074070 -0.20254325 -0.01738603
##  [6,]  1.62824362 -0.07145082  0.08968766 -0.31716549 -0.19076604
##  [7,] -0.07145082  1.67496231 -0.39533832 -0.20238905 -0.28629921
##  [8,]  0.08968766 -0.39533832  1.29540846  0.06472187 -0.15643134
##  [9,] -0.31716549 -0.20238905  0.06472187  1.43858106 -0.44178223
## [10,] -0.19076604 -0.28629921 -0.15643134 -0.44178223  1.58086841
```

```
# find the variance for the 1st and 3rd residuals
evar[1,1]
```

```
## [1] 1.316208
```

```
evar[3,3]
```

```
## [1] 1.295119
```

So the variance for the 1st residual is 1.316208, and the variance for the 3rd residual is 1.295119.

d).

```
# find the covariance between the first and third residual
evar[1,3]
```

```
## [1] -0.4239149
```

So the covariance between the first and third residual is -0.4239149.

3.

a).

```
#read in data from Center for Radiative Shock Hydrodynamics (crash)
crash=read.csv("data_computer_experiment.csv")

#attach that dataframe to use variable names in the input file
attach(crash)

#fit linear model and summarize
crash.lm <- lm(location~thickness+energy+flux+gamma+opacity)
(crash.sum <- summary(crash.lm))
```

```
##
## Call:
## lm(formula = location ~ thickness + energy + flux + gamma + opacity)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.861  -4.121  -0.088   4.677  20.658
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.503e+02  3.058e+01  24.533  < 2e-16 ***
## thickness    1.739e+01  6.324e-01  27.492  < 2e-16 ***
## energy      -2.344e-02  6.777e-03  -3.459 0.000817 ***
## flux        -3.360e+02  7.344e+00 -45.750  < 2e-16 ***
## gamma        1.055e+00  4.506e+00   0.234 0.815484
## opacity     -1.251e+03  1.026e+02 -12.196  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.245 on 94 degrees of freedom
## Multiple R-squared:  0.9726, Adjusted R-squared:  0.9711
## F-statistic: 666.5 on 5 and 94 DF,  p-value: < 2.2e-16
```

```
# compute the the estimated variance for the least squares estimator of the regression
#coefficient for the thickness of the beryllium disk
crash.sum$coefficients[2,2]^2
```

```
## [1] 0.3999662
```

So the estimated variance for the least squares estimator of the regression coefficient for the thickness of the beryllium disk is 0.3999662.

b).

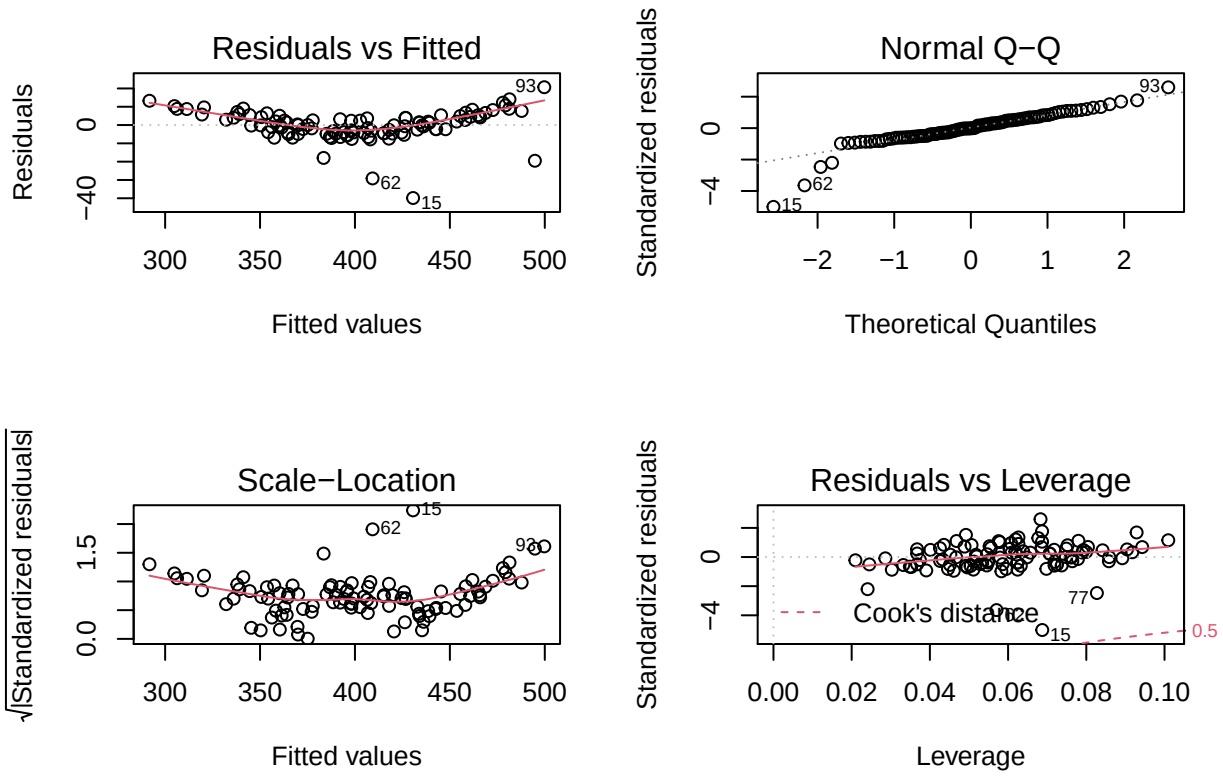
```
# Combine X matrix
X <- cbind(rep(1,100),thickness,energy,flux,gamma,opacity)

# Compute the covariance for the least squares regression estimators
solve(t(X) %*% X) * crash.sum$sigma^2
```

```
##                           thickness        energy          flux         gamma
##            935.2395345 -8.0088121283 -1.637646e-01  -62.60013306 -19.317017440
## thickness   -8.0088121  0.3999662172 -1.055582e-04    0.04179892   0.132474492
## energy      -0.1637646 -0.0001055582  4.592843e-05   -0.00439205  -0.001119203
## flux       -62.6001331  0.0417989235 -4.392050e-03   53.93636554  -0.285660947
## gamma      -19.3170174  0.1324744921 -1.119203e-03   -0.28566095  20.308181978
## opacity   -509.0163943  3.1840600198 -1.783864e-02 -102.59370962  18.190484520
##                opacity
##           -5.090164e+02
## thickness  3.184060e+00
## energy    -1.783864e-02
## flux      -1.025937e+02
## gamma      1.819048e+01
## opacity    1.052006e+04
```

So the estimated covariance between the least squares estimators of the regression coefficients for the thickness of the beryllium disk and the wall opacity is 3.1840600198.

c).

```
par(mfrow = c(2, 2))
plot(crash.lm)
```



i Check the constant variance assumption for the error

According to the Residuals vs Fitted plot, we can see a "smiley face", the residual goes down before fitted value is 400, and goes up after fitted value is 400. However it is not too bad. Most of the points are still within the horizontal band(let's say residuals from -10 to 20). Moreover according to the Scale-Location plot, it is almost a straight line and residuals are spread equally above and below it. So i would conclude that the constant variance assumption is satisfied.

ii Check the normality assumption.

The Normal Q-Q plot shows almost all the points fall on the straight dotted line except a few point on the lower tail, so the normality of errors assumption is satisfied.

iii Check for large leverage points.

According to Residuals vs leverage plot, we see none of the points in our dataset has both high leverage and high standardized residuals.

iv Check for outliers.

```
# Compute studentized residuals
rstudent(crash.lm)
```

```
##             1             2             3             4             5
## -9.426509e-01 -3.285465e-01  1.552826e-01 -5.374069e-01  9.629384e-01
##             6             7             8             9            10
##  1.531708e+00  8.238701e-01  1.905079e-01 -3.697197e-02 -2.209180e-01
##            11            12            13            14            15
##  7.425703e-01 -2.020077e-01 -3.717327e-01 -4.817109e-01 -5.820794e+00
##            16            17            18            19            20
## -5.040435e-01  1.584154e-01 -5.459220e-01 -8.097802e-01 -5.757016e-01
##            21            22            23            24            25
##  1.051894e+00 -8.758435e-02  1.101707e+00 -2.696817e-01  4.381856e-02
##            26            27            28            29            30
## -1.369282e-01 -2.696825e-01 -2.184272e-02 -3.853124e-01  3.044765e-01
##            31            32            33            34            35
## -6.833062e-01  3.969862e-01 -5.770374e-01 -2.227737e-02  4.962796e-01
##            36            37            38            39            40
## -4.041027e-01  8.350949e-01  6.755291e-01 -8.171225e-01  6.094770e-01
##            41            42            43            44            45
## -6.112150e-01 -8.719479e-01  7.131934e-01 -4.892700e-01  6.310281e-01
##            46            47            48            49            50
##  2.404224e-01  5.311699e-01 -4.950943e-01 -8.183808e-01 -7.102644e-01
##            51            52            53            54            55
## -5.984827e-01 -8.638650e-01  1.789942e+00 -2.880734e-01 -9.199789e-01
##            56            57            58            59            60
##  1.023971e+00 -6.041084e-01  1.212145e+00  4.887678e-01  1.094989e+00
##            61            62            63            64            65
## -2.253199e+00 -3.908698e+00 -1.644429e-01 -3.774412e-01 -5.890033e-01
##            66            67            68            69            70
## -8.617846e-01  5.219459e-01  1.705509e+00  2.359979e-01  6.900922e-01
##            71            72            73            74            75
## -2.614203e-02  3.003619e-01 -9.797174e-01  5.573971e-01  1.155427e+00
##            76            77            78            79            80
## -6.589834e-01 -2.541339e+00  8.073505e-01  4.618926e-01 -2.898771e-01
##            81            82            83            84            85
##  8.974194e-01  6.810672e-01  1.736733e-01 -5.171131e-01  3.477007e-01
##            86            87            88            89            90
##  3.641381e-01  1.351680e+00  1.069454e-05  1.304322e+00 -5.611701e-03
##            91            92            93            94            95
##  4.824116e-01  3.267218e-01  2.679931e+00  5.770485e-01 -8.286678e-02
##            96            97            98            99           100
##  2.325489e-01 -3.119475e-01 -1.661301e-02  1.117339e+00  2.944201e-01
```

According to the plot and studentized residuals, point 15 and point 62 have studentized residuals greater than 3, so these two points are potential outliers.

```
# Construct new dataset without point 15 and point 62.
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------------------
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
```

```
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------ t
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
newcrash <- filter(crash,thickness!= 22.00 & thickness != 21.17)

# fit linear model with new dataset
newcrash.lm <- lm(data=newcrash,location~thickness+energy+flux+gamma+opacity)

# Compare with old linear model
crash.sum
```

```
##
## Call:
## lm(formula = location ~ thickness + energy + flux + gamma + opacity)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.861  -4.121  -0.088   4.677  20.658
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.503e+02  3.058e+01  24.533  < 2e-16 ***
## thickness    1.739e+01  6.324e-01  27.492  < 2e-16 ***
## energy      -2.344e-02  6.777e-03  -3.459 0.000817 ***
## flux        -3.360e+02  7.344e+00 -45.750  < 2e-16 ***
## gamma        1.055e+00  4.506e+00   0.234 0.815484
## opacity     -1.251e+03  1.026e+02 -12.196  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.245 on 94 degrees of freedom
## Multiple R-squared:  0.9726, Adjusted R-squared:  0.9711
## F-statistic: 666.5 on 5 and 94 DF,  p-value: < 2.2e-16
```

```
summary(newcrash.lm)
```

```
##
## Call:
## lm(formula = location ~ thickness + energy + flux + gamma + opacity,
##     data = newcrash)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.2049  -4.2077  -0.3152   3.9673  18.8923
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.400e+02  2.404e+01  30.788  < 2e-16 ***
## thickness    1.806e+01  5.008e-01  36.070  < 2e-16 ***
## energy      -2.424e-02  5.382e-03  -4.505 1.95e-05 ***
## flux        -3.338e+02  5.737e+00 -58.187  < 2e-16 ***
## gamma       -6.637e-01  3.586e+00  -0.185    0.854
```

```
## opacity     -1.271e+03  8.033e+01 -15.826  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.432 on 92 degrees of freedom
## Multiple R-squared:  0.9836, Adjusted R-squared:  0.9827
## F-statistic:  1105 on 5 and 92 DF,  p-value: < 2.2e-16
```

Deleting point 15 and 62 has almost no effect on the estimates of the regression coefficients except gamma; however, i noticed that the p-value for gamma is way larger than significant level 0.05, so we fail to reject the hypothesis that beta(gamma) is 0 in presence of thickness, energy, flux, and opacity in the model, and since point 15 and point 62 are considering not affect other coefficients except gamma, so i would conclude that point 15 and 62 are not outliers.

v. Check for influential points.

```
# hat matrix
H <- X %*% solve(t(X) %*% X) %*% t(X)
# hii
diag(H)
```

```
##   [1] 0.04611532 0.07190267 0.05177169 0.04513030 0.06075931 0.04921654
##   [7] 0.05065548 0.05462249 0.05795402 0.02095385 0.05528178 0.03745519
##  [13] 0.05487036 0.05287465 0.06873617 0.02452466 0.04523320 0.03320825
##  [19] 0.06979128 0.07370306 0.06868700 0.02860796 0.04681324 0.04261557
##  [25] 0.05194017 0.07786541 0.05483530 0.06482732 0.06232485 0.07899337
##  [31] 0.03522345 0.06382778 0.05453865 0.04671679 0.07178507 0.05336727
##  [37] 0.04358395 0.06857212 0.04463337 0.04257254 0.05041226 0.03026844
##  [43] 0.04885847 0.05859770 0.05562216 0.03675337 0.03672842 0.03673035
##  [49] 0.05952433 0.04956529 0.06324849 0.06243377 0.06869074 0.07641636
##  [55] 0.03907708 0.05787671 0.04919138 0.06191795 0.04014103 0.05758462
##  [61] 0.02407434 0.05704505 0.04325205 0.06249178 0.07189441 0.05112985
##  [67] 0.09041658 0.09286574 0.07069150 0.08042332 0.07600100 0.06569162
##  [73] 0.05829172 0.06038033 0.10095996 0.03487529 0.08270483 0.07521326
##  [79] 0.08415731 0.08577572 0.06265067 0.09429105 0.06217456 0.07251430
##  [85] 0.07996523 0.06186785 0.06265332 0.07479033 0.06772879 0.08586913
##  [91] 0.07925925 0.09176931 0.06829407 0.07489126 0.08920297 0.07061004
##  [97] 0.05127946 0.07626523 0.07798380 0.06087337
```

According to residual vs leverage plot, we do not see any points appear outside of cooks distance line marked by the dotted red line; moreover, we do not observe any high hii points.

vi. Check the structure of the relationship between the predictors and the response.

We discussed about gamma in previous question, we will discuss it further in this question.

```
# check Multicollinearity
library(faraway)
vif(crash.lm)
```

```
## thickness    energy      flux     gamma   opacity
##  1.005095  1.011063  1.027967  1.004812  1.024190
```

since VIF is not greater than 10, we do not need to worry about multicollinearity

```
# Partial regression plot
library(car)
```

```
## Loading required package: carData
```

```
## Registered S3 methods overwritten by 'car':
##   method                           from
##   influence.merMod                 lme4
##   cooks.distance.influence.merMod lme4
##   dfbeta.influence.merMod          lme4
##   dfbetas.influence.merMod         lme4
```

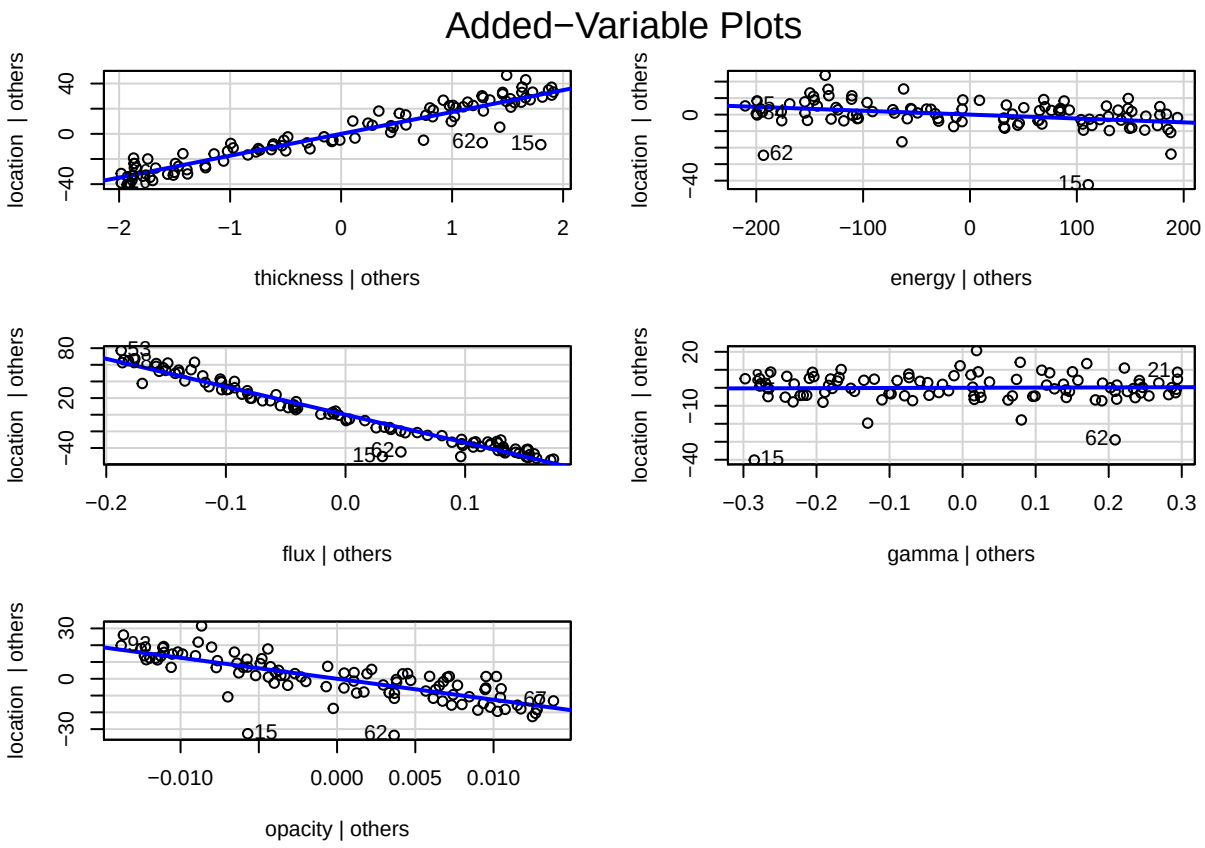
```
##
## Attaching package: 'car'
```

```
## The following objects are masked from 'package:faraway':
##
##     logit, vif
```

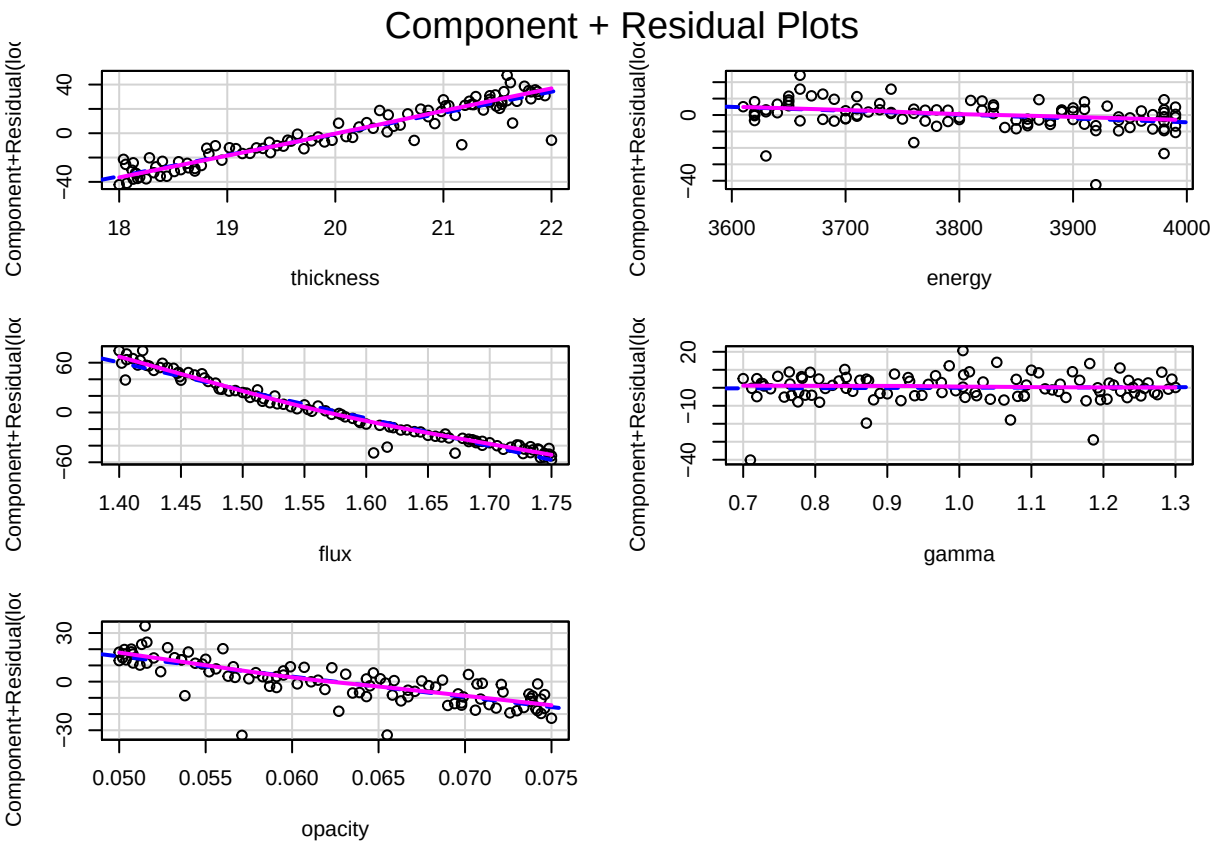
```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
avPlots(crash.lm)
```




```
# Partial residual plot
crPlots(crash.lm)
```



If predict variables have an linear relationship with response variables, then these plots should display a linear pattern; however, we do not see a clear linear pattern between gamma and location, and we discussed hypothesis test in previous question, so we can further conclude that there is not an linear relationship between location and gamma. Further analyses (transformation) are needed to answer the question whether we should include gamma variable in our model.