Homework 2

Netid: ghosh17

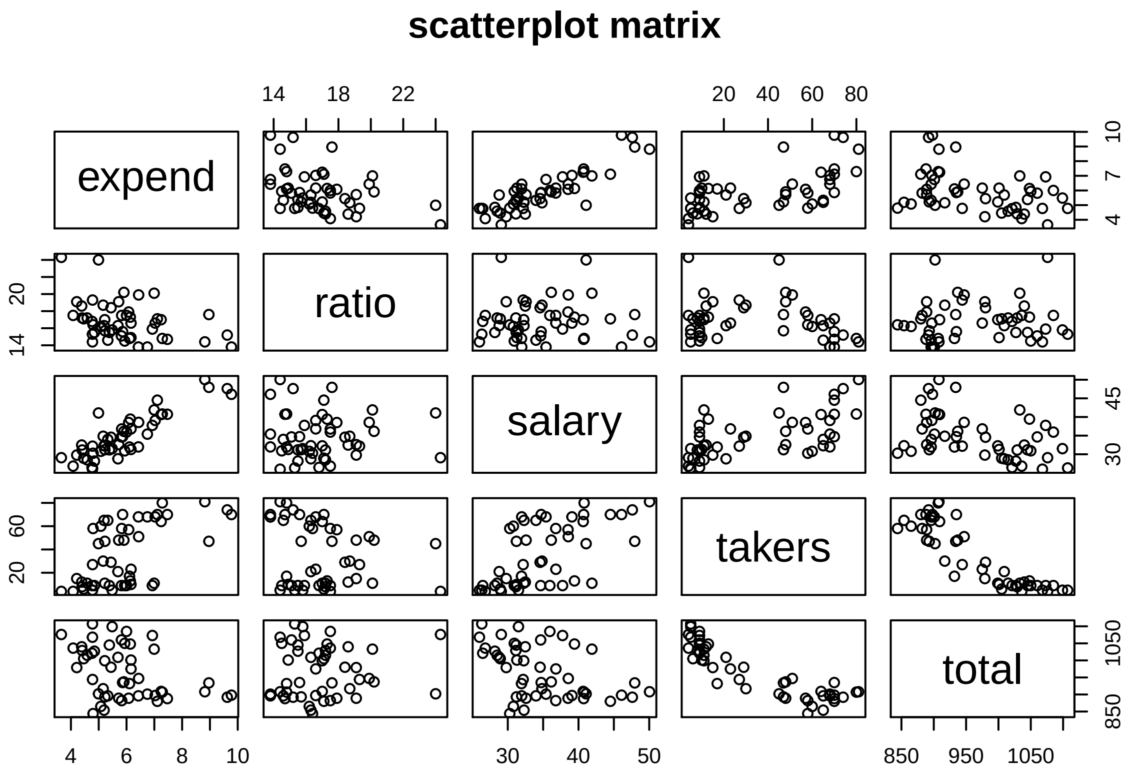
Subhankar Ghosh

Question 1

```
library(faraway)
data(sat)
satdata = sat
#Looking at the data
head(satdata)
```

```
##            expend ratio salary takers verbal math total
## Alabama     4.405  17.2 31.144      8    491  538  1029
## Alaska      8.963  17.6 47.951     47    445  489   934
## Arizona     4.778  19.3 32.175     27    448  496   944
## Arkansas    4.459  17.1 28.934      6    482  523  1005
## California  4.992  24.0 41.078     45    417  485   902
## Colorado    5.443  18.4 34.571     29    462  518   980
```

```
# Looking at the scatterplot of all data
pairs(~expend + ratio + salary + takers + total, data = satdata, main = "scatterplot matrix")
```



```r
model = lm(total ~ expend + ratio + salary + takers, data = satdata)
summary(model)
```

```
##
## Call:
## lm(formula = total ~ expend + ratio + salary + takers, data = satdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -90.531 -20.855  -1.746  15.979  66.571
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1045.9715    52.8698  19.784  < 2e-16 ***
## expend         4.4626    10.5465   0.423    0.674
## ratio         -3.6242     3.2154  -1.127    0.266
## salary         1.6379     2.3872   0.686    0.496
## takers        -2.9045     0.2313 -12.559 2.61e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.7 on 45 degrees of freedom
## Multiple R-squared:  0.8246, Adjusted R-squared:  0.809
## F-statistic: 52.88 on 4 and 45 DF,  p-value: < 2.2e-16
```

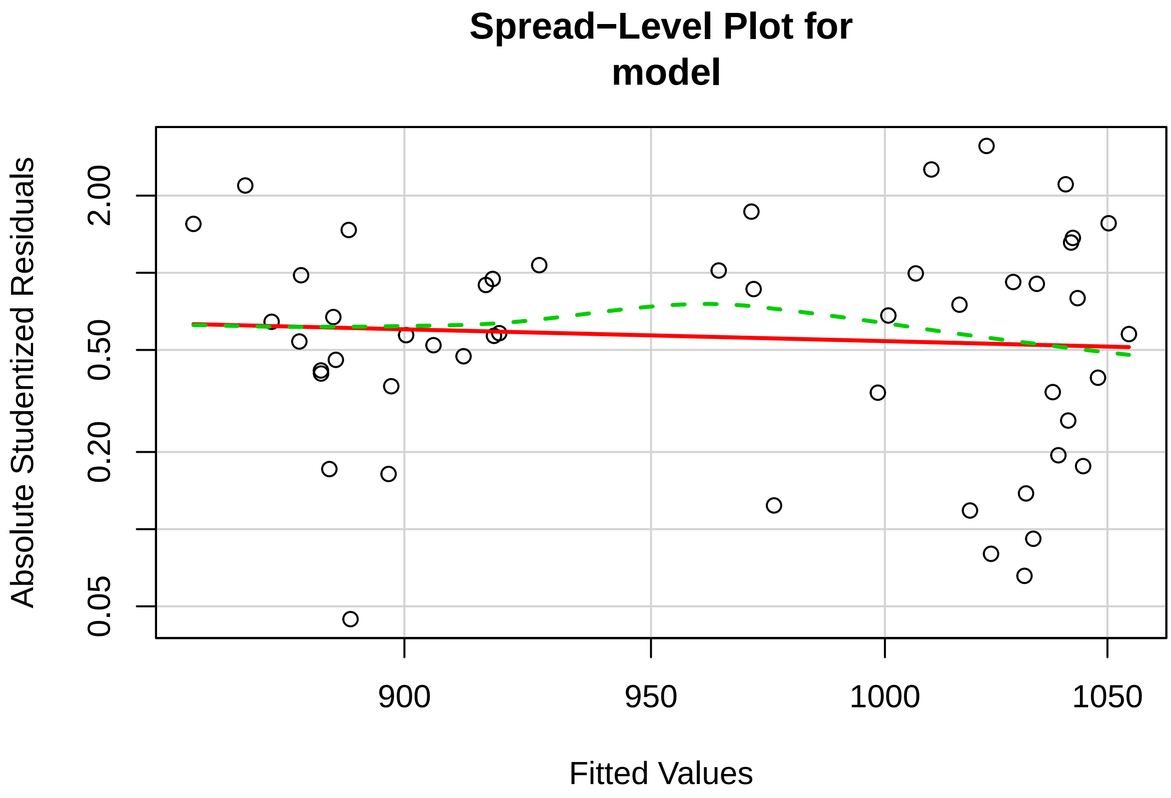
a. Check the constant variance assumption.

```r
car::ncvTest(model)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.6972119    Df = 1     p = 0.4037221
```

In this hypothesis testing the H_0 : Variance is constant and since $p - value = 0.403$ is greater that $\alpha = 0.05$ we can accept the null hypothesis that the *Variance is actually constant.*

```r
car::spreadLevelPlot(model)
```

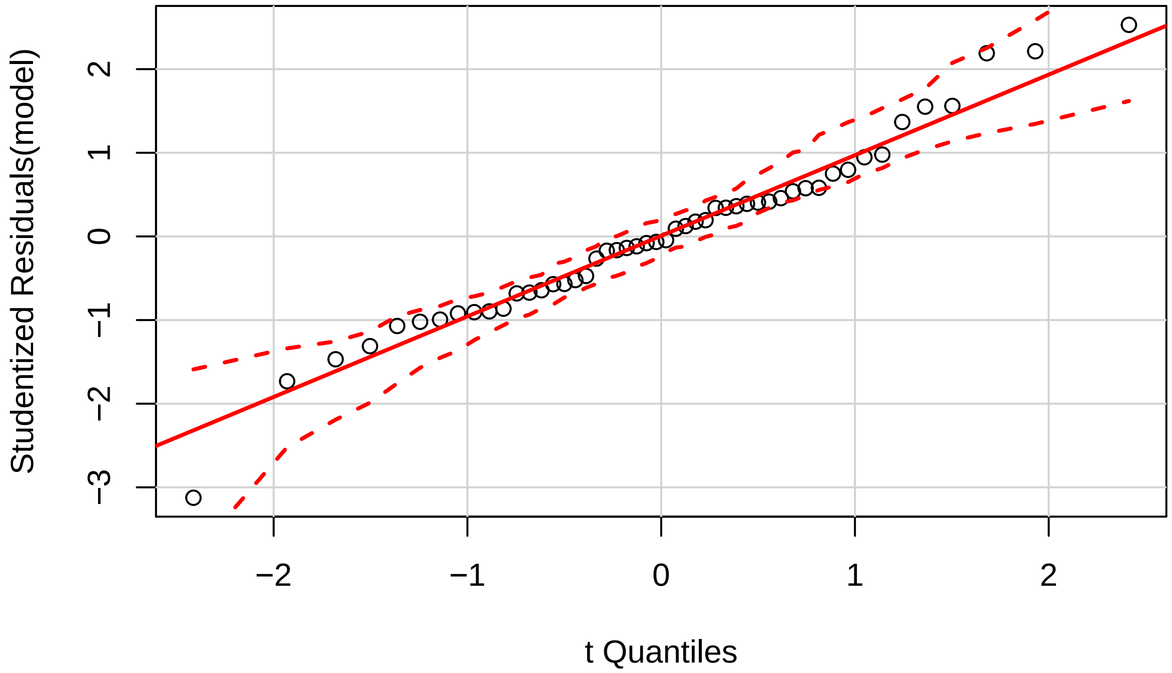


```
##
## Suggested power transformation:  2.006005
```

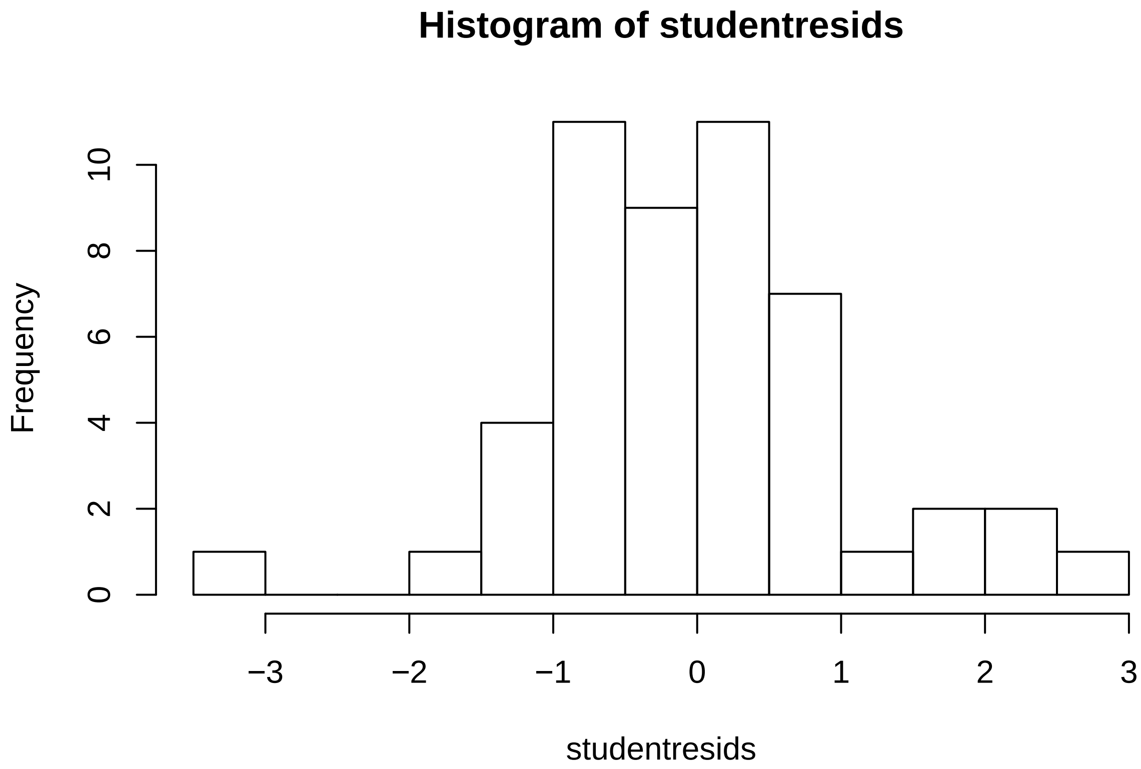
There is very slight evidence of variance going down with the mean but it is not statistically significant.

b. Check the normality assumption.

```
car::qqPlot(model)
```



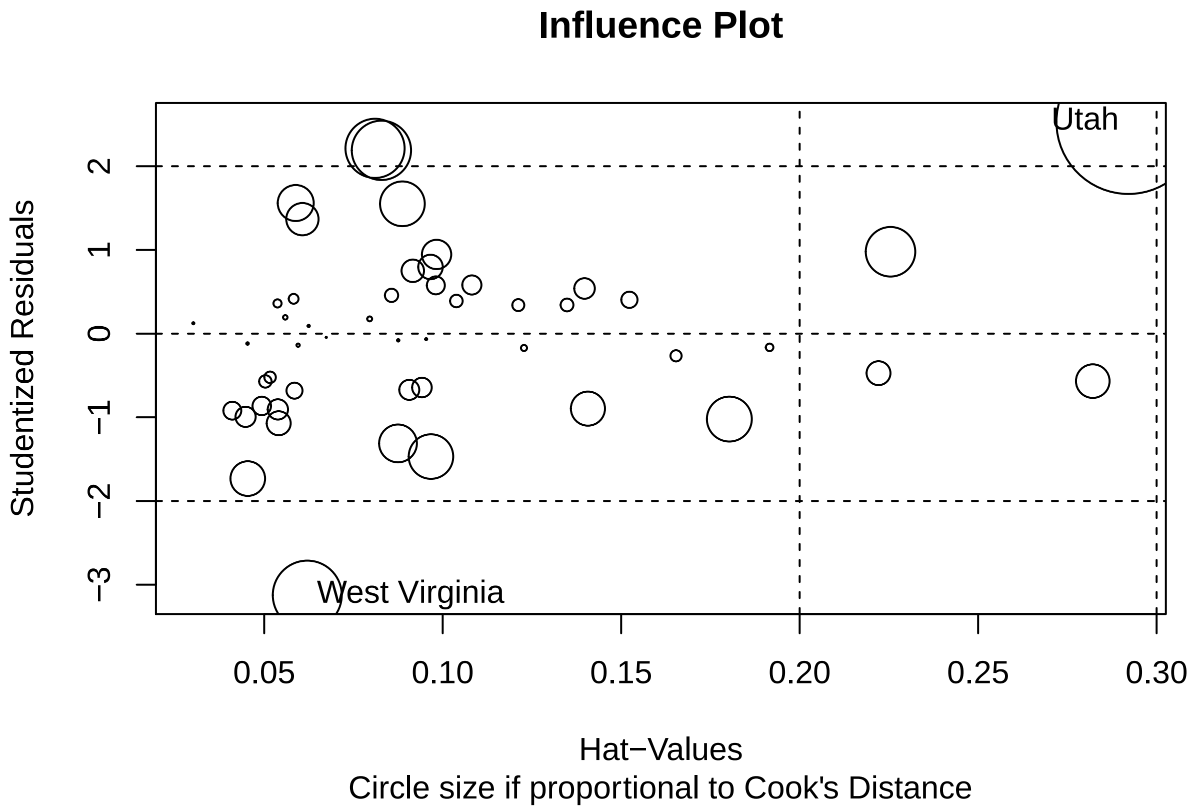
```
studentresids=rstudent(model)
hist(studentresids,nclass=12)
```



From the plot we can conclude that the residuals are normally distributed.

c. Check for large leverage points.

```
car::influencePlot(model, main="Influence Plot", sub="Circle size if proportional to Cook's Distance")
```



```
##                  StudRes        Hat     CookD
## Utah            2.529587 0.29211280 0.4715287
## West Virginia -3.124428 0.06206536 0.1081395
```

West Virginia and Utah have quite large leverage points.

d) Check for outliers.

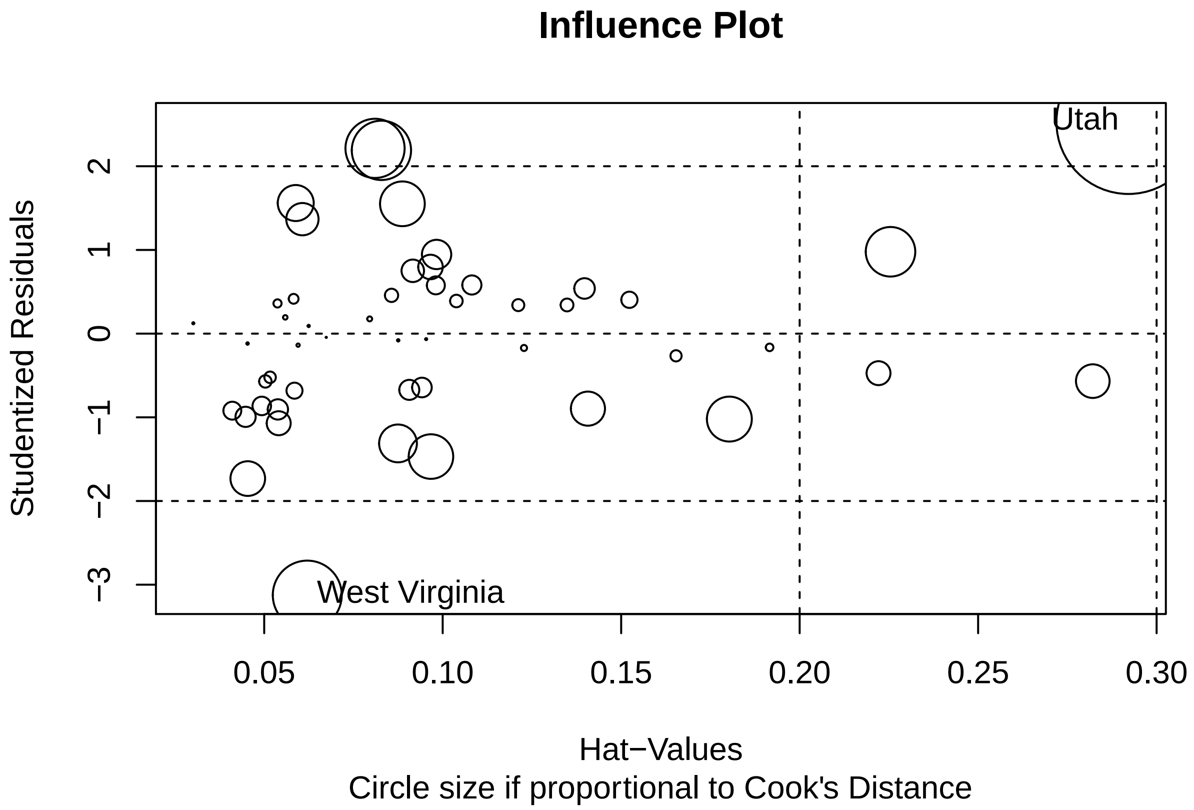
```
car::outlierTest(model)
```

```
##
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##                 rstudent unadjusted p-value Bonferonni p
## West Virginia -3.124428          0.0031496      0.15748
```

There appears to be no evidence of outliers

e. Check for influential points.

```
car::influencePlot(model, main="Influence Plot", sub="Circle size if proportional to Cook's Distance")
```

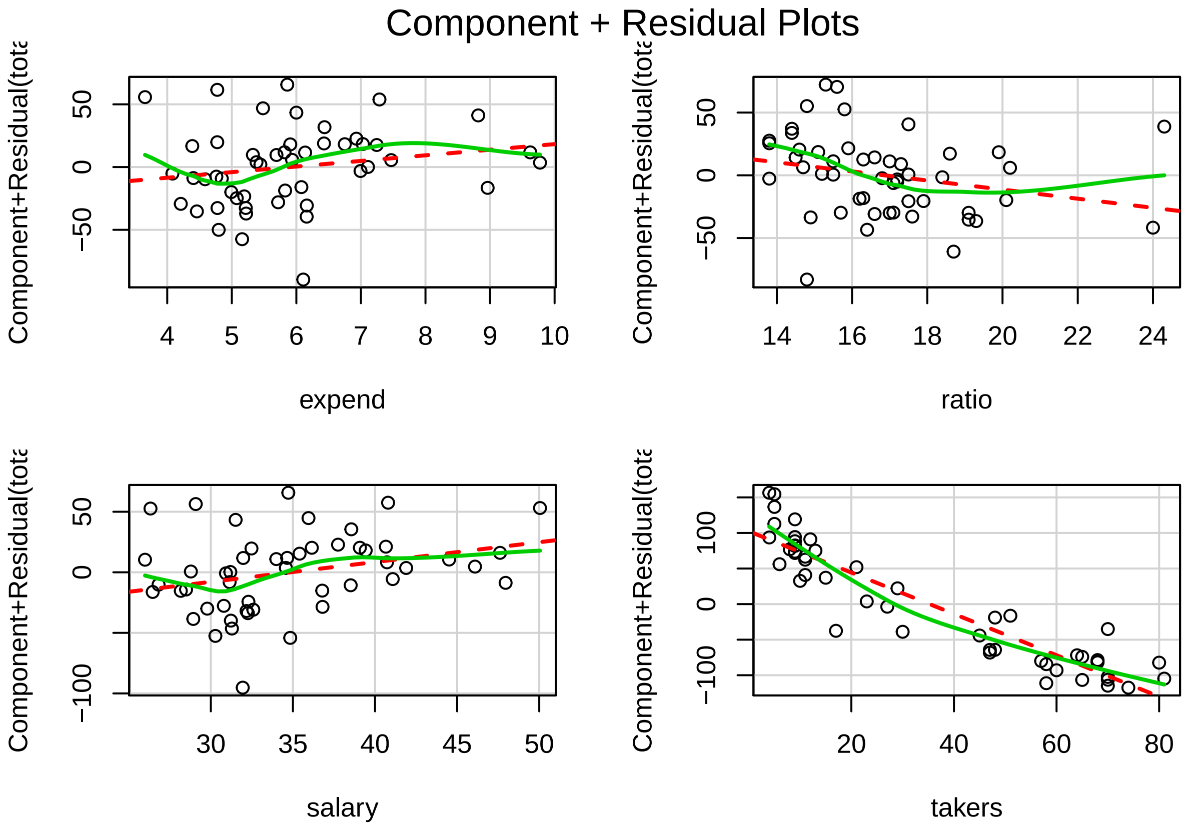


```
##                   StudRes        Hat      CookD
## Utah             2.529587 0.29211280 0.4715287
## West Virginia -3.124428 0.06206536 0.1081395
```

West Virginia and Utah are quite influential points.

f. Check the functional form of the relationship between the predictors and the response.

```
car::crPlots(model)
```



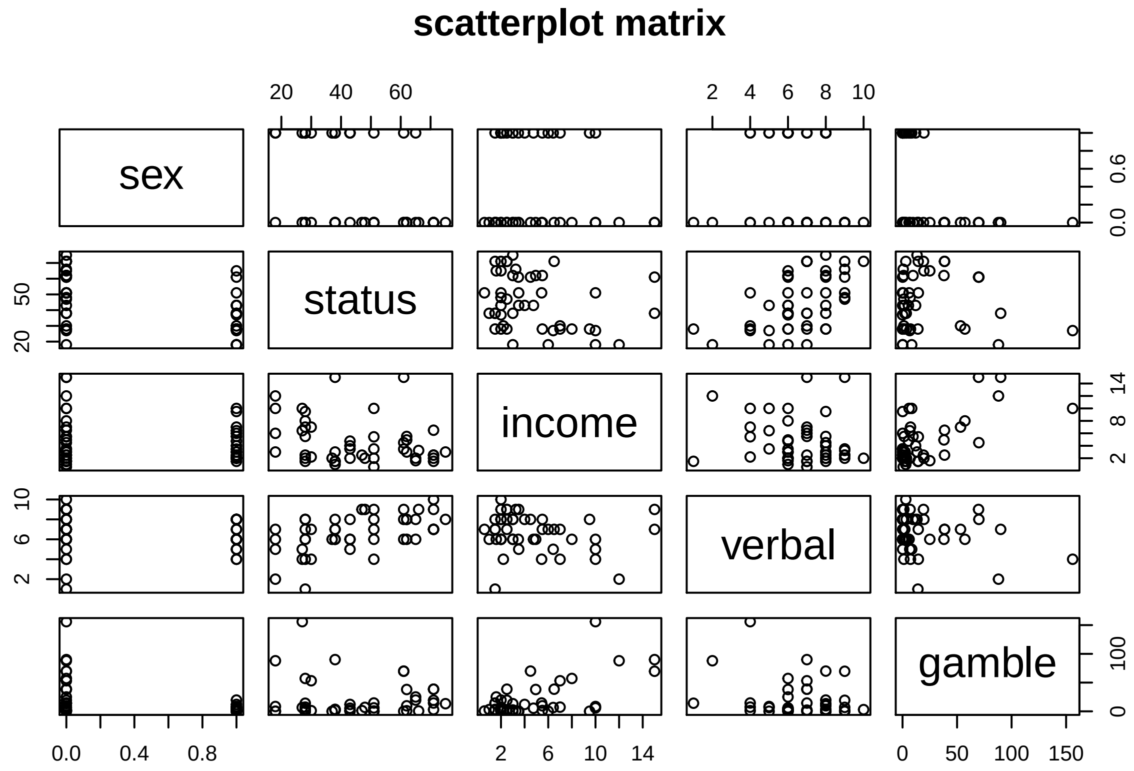
From the functional form plots we can notice that there is slight curvature in all the predictors (expend, ratio, salary, takers)

Question 2

```r
data("teengamb")
teengamb = teengamb
#Looking at the data
head(teengamb)
```

```
##   sex status income verbal gamble
## 1   1     51   2.00      8    0.0
## 2   1     28   2.50      8    0.0
## 3   1     37   2.00      6    0.0
## 4   1     28   7.00      4    7.3
## 5   1     65   2.00      8   19.6
## 6   1     61   3.47      6    0.1
```

```r
# Looking at the scatterplot of all data
pairs(~sex + status + income + verbal + gamble, data = teengamb, main = "scatterplot matrix")
```



```
model1 = lm(gamble ~ sex + status + income + verbal, data = teengamb)
summary(model1)
```

```
##
## Call:
## lm(formula = gamble ~ sex + status + income + verbal, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.082 -11.320  -1.451   9.452  94.252
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.55565   17.19680   1.312   0.1968
## sex         -22.11833    8.21111  -2.694   0.0101 *
## status        0.05223    0.28111   0.186   0.8535
## income        4.96198    1.02539   4.839 1.79e-05 ***
## verbal       -2.95949    2.17215  -1.362   0.1803
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.69 on 42 degrees of freedom
## Multiple R-squared:  0.5267, Adjusted R-squared:  0.4816
## F-statistic: 11.69 on 4 and 42 DF,  p-value: 1.815e-06
```

a. Check the constant variance assumption.

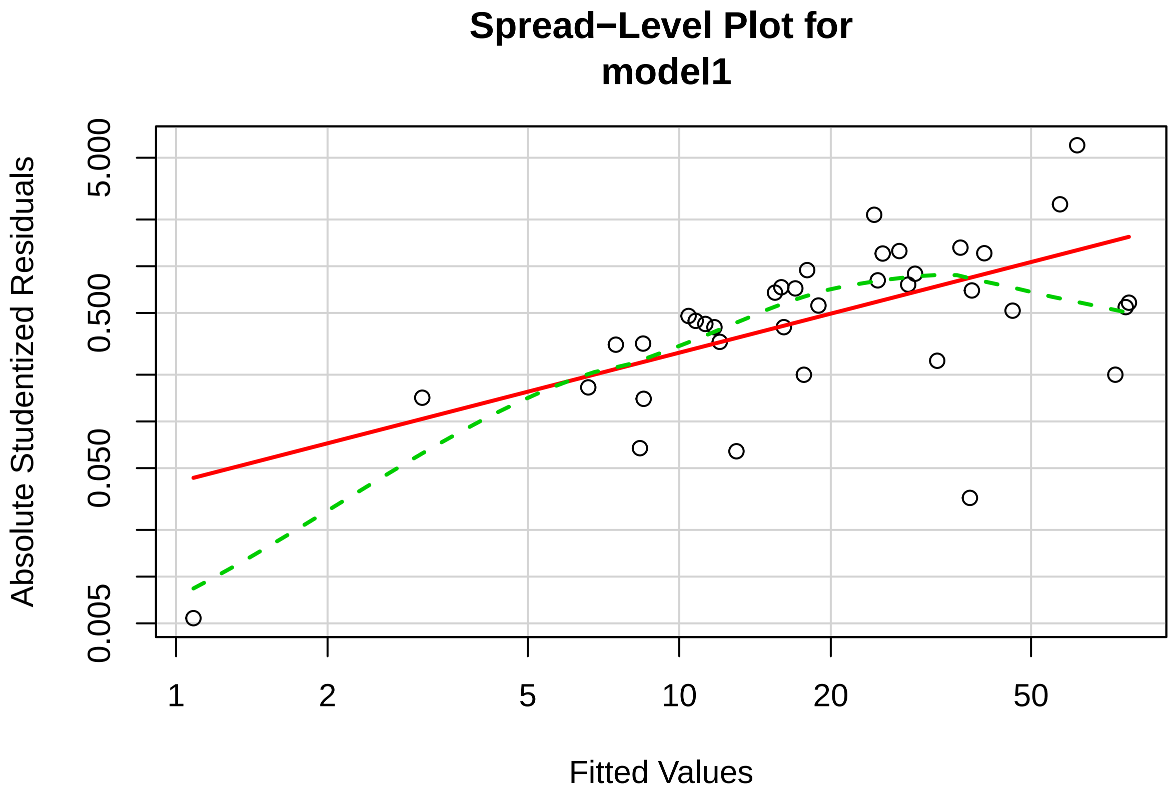
```
car::ncvTest(model1)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 24.29051    Df = 1     p = 8.284638e-07
```

In this hypothesis testing the H_0 : Variance is constant and since $p - value$ is less than $\alpha = 0.05$ we cannot accept the null hypothesis. That is the *Variance is actually not constant.*

```
car::spreadLevelPlot(model1)
```

```
## Warning in spreadLevelPlot.lm(model1): 10 negative fitted values removed
```

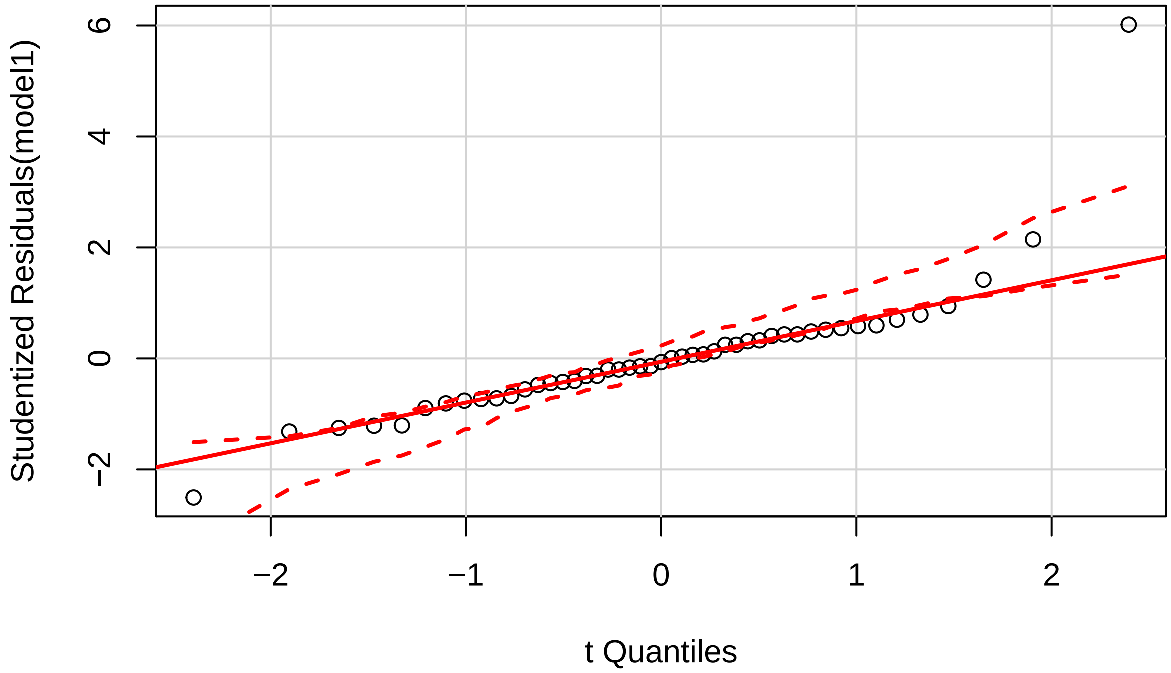


```
##
## Suggested power transformation:  0.1646836
```

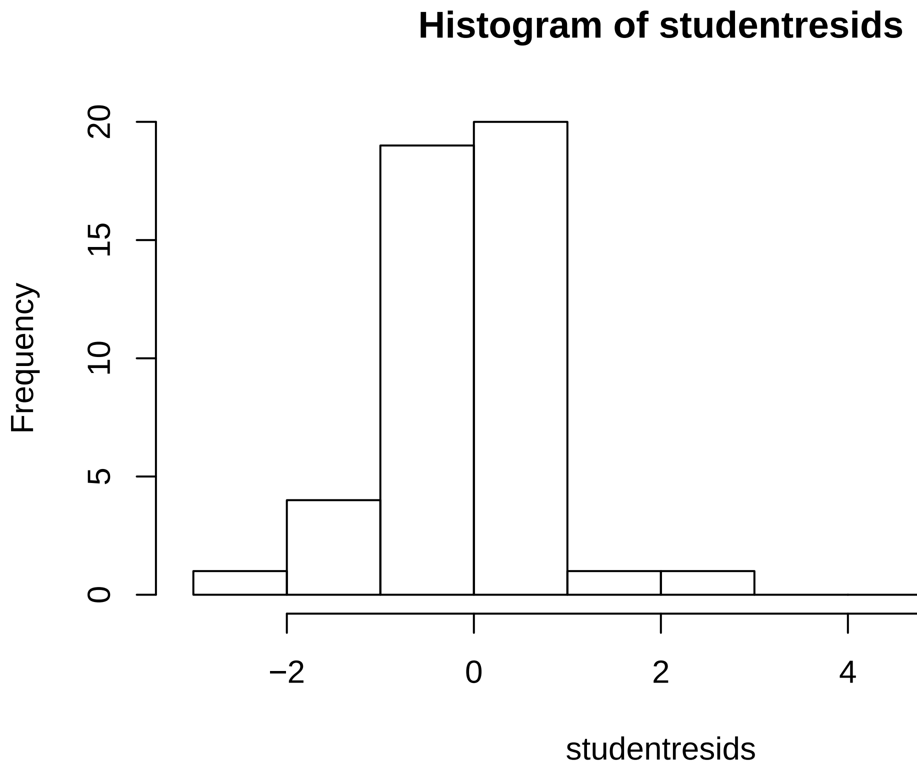
There is strong evidence of variance going up with the mean.

b. Check the normality assumption.

```
car::qqPlot(model1)
```



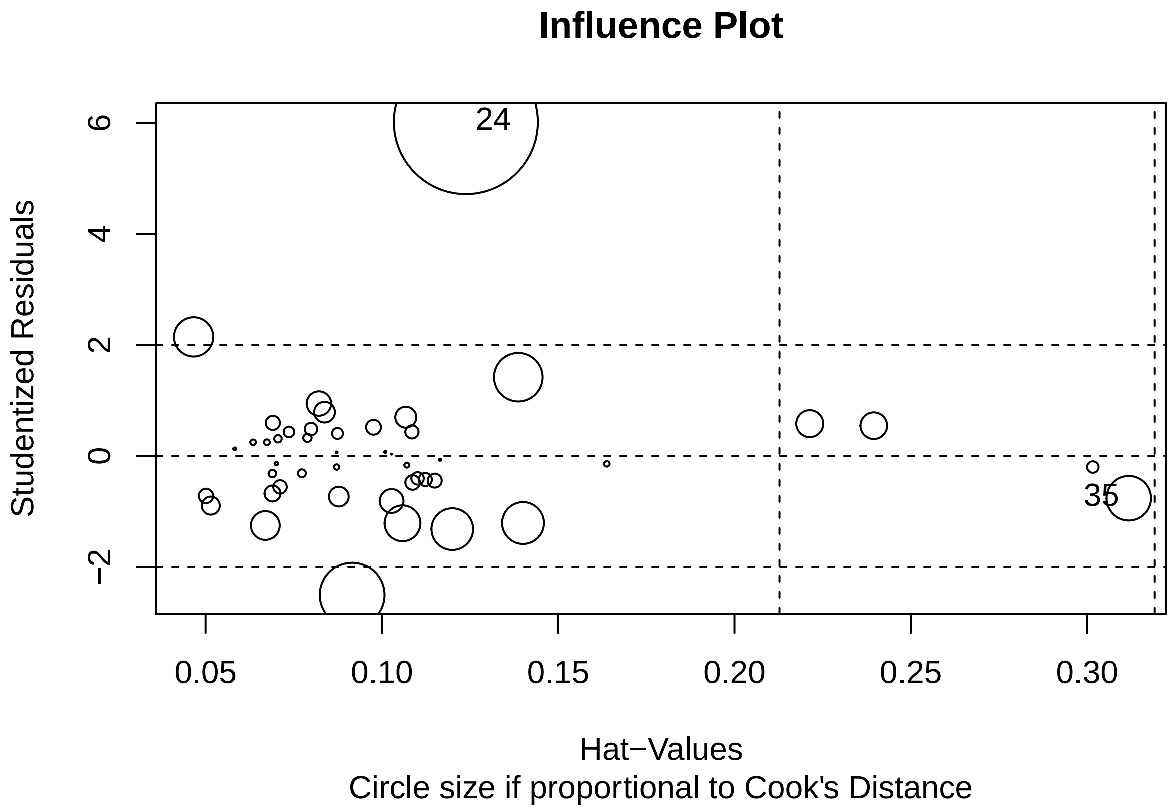
```
studentresids=rstudent(model1)
hist(studentresids,nclass=10)
```



From the plot we can conclude that the residuals are not normally distributed as the histogram does not correspond to bell shape and qqPlot does not correspond to a straight line.

c. Check for large leverage points.

```
car::influencePlot(model1, main="Influence Plot", sub="Circle size if proportional to Cook's Distance")
```



```
##        StudRes       Hat      CookD
## 24  6.0161163 0.1238046 0.55650113
## 35 -0.7612557 0.3118029 0.05304304
```

24 and 35 have quite large leverage points.

d) Check for outliers.

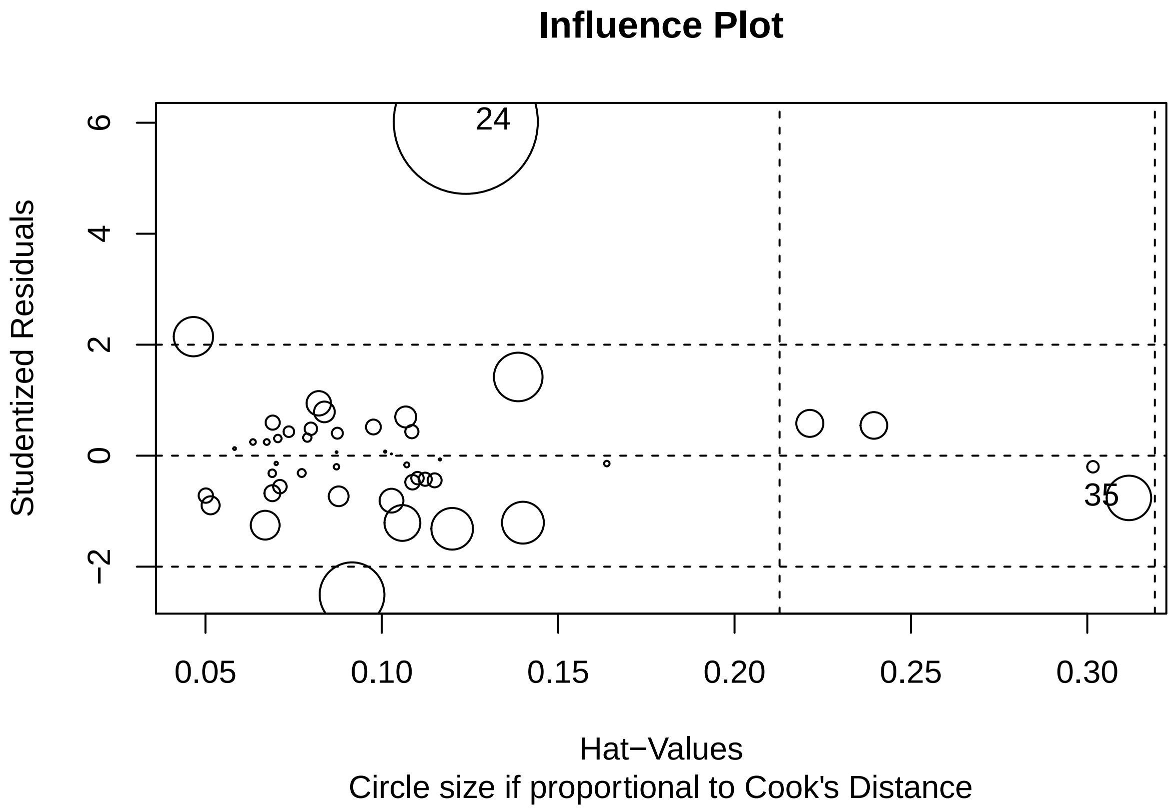
```
car::outlierTest(model1)
```

```
##    rstudent unadjusted p-value Bonferonni p
## 24 6.016116         4.1041e-07   1.9289e-05
```

There appears to be strong evidence of outliers.

e. Check for influential points.

```
car::influencePlot(model1, main="Influence Plot", sub="Circle size if proportional to Cook's Distance")
```

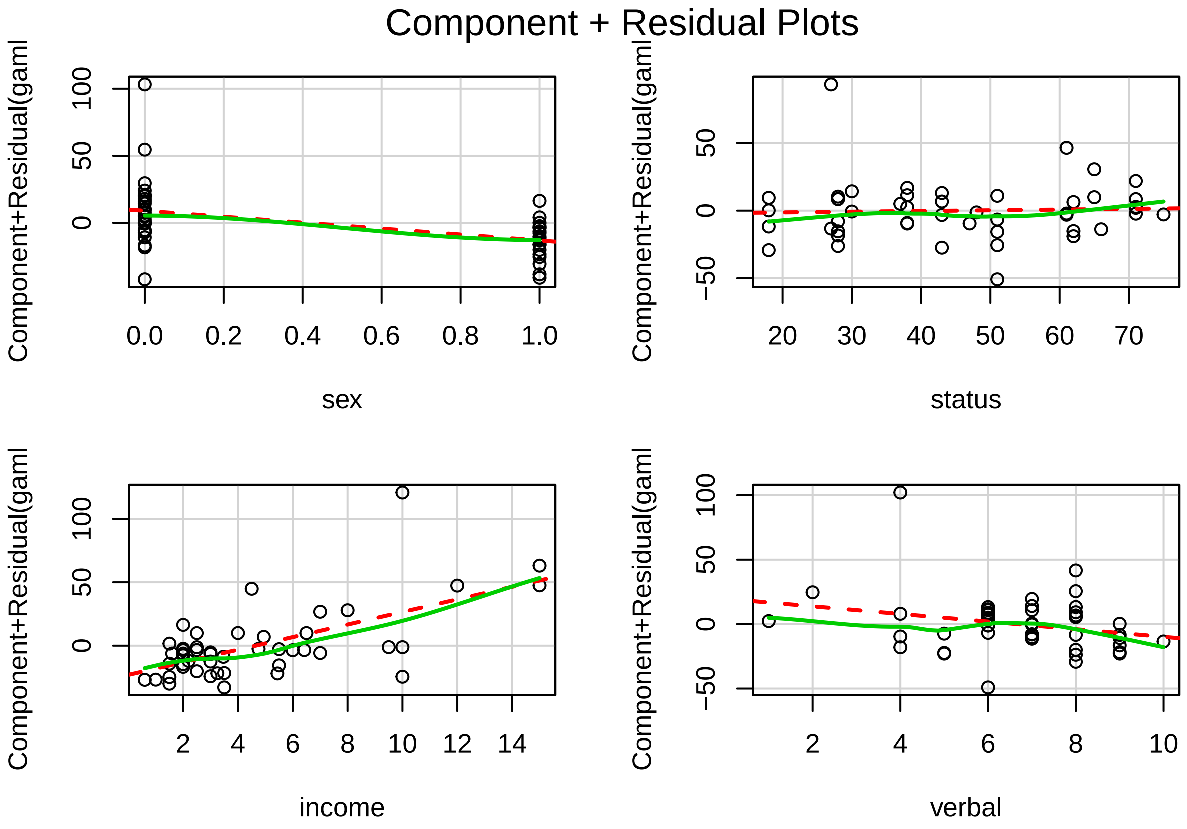


```
##        StudRes       Hat      CookD
## 24  6.0161163 0.1238046 0.55650113
## 35 -0.7612557 0.3118029 0.05304304
```

24 and 35 seem to be quite influencial points.

f. Check the functional form of the relationship between the predictors and the response.

```
car::crPlots(model1)
```



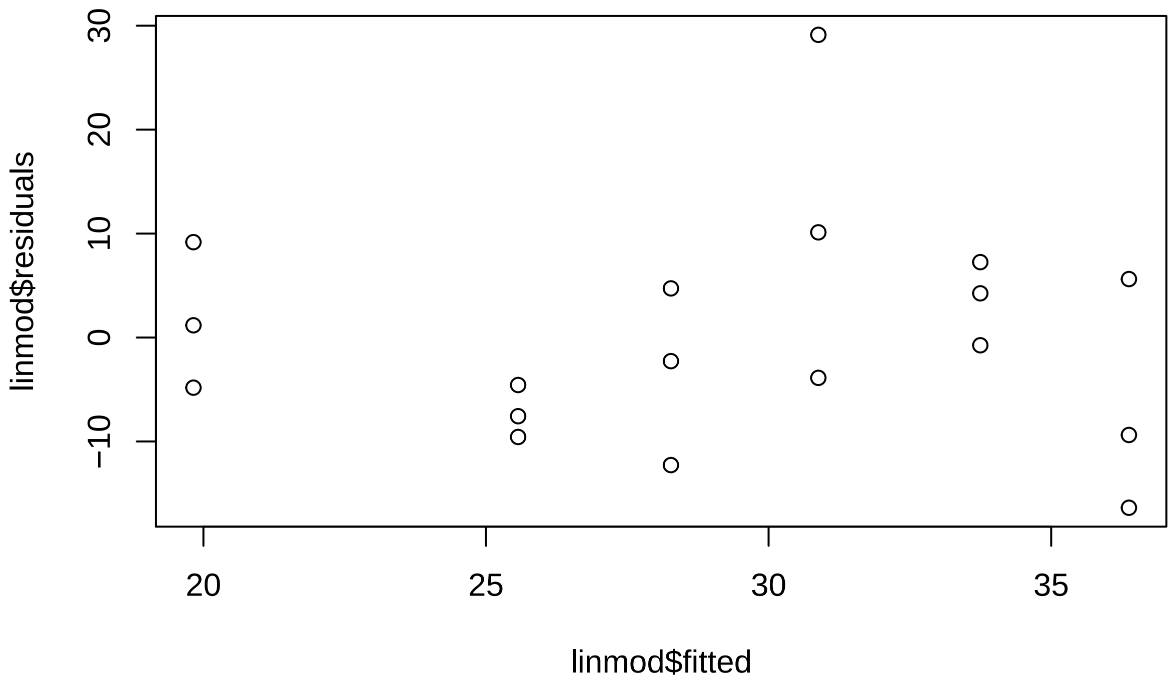
From the functional form plots we can notice that there is slight curvature in verbal and income predictors.

Question 3

```
head(salmonella)
```

```
##   colonies dose
## 1       15    0
## 2       21    0
## 3       29    0
## 4       16   10
## 5       18   10
## 6       21   10
```

```
linmod=lm(colonies~log(dose + 1), data = salmonella)
plot(linmod$fitted,linmod$residuals)
```



The lack of constant variance is quite evident from the residual vs fitted plot but let is run a test

```
car::ncvTest(linmod)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.825563    Df = 1     p = 0.3635587
```

```
genmod=lm(colonies~factor(log(dose + 1)), data = salmonella)
summary(genmod)
```

```
##
## Call:
## lm(formula = colonies ~ factor(log(dose + 1)), data = salmonella)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.667  -3.917  -0.500   3.417  17.333
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           21.667      5.506   3.935  0.00198
## factor(log(dose + 1))2.39789527279837   -3.333      7.787  -0.428  0.67617
## factor(log(dose + 1))3.52636052461616    3.333      7.787   0.428  0.67617
## factor(log(dose + 1))4.61512051684126   21.000      7.787   2.697  0.01942
## factor(log(dose + 1))5.8111409929767    15.667      7.787   2.012  0.06722
## factor(log(dose + 1))6.90875477931522    8.000      7.787   1.027  0.32449
##
```

```
## (Intercept)                           **
## factor(log(dose + 1))2.39789527279837
## factor(log(dose + 1))3.52636052461616
## factor(log(dose + 1))4.61512051684126 *
## factor(log(dose + 1))5.8111409929767  .
## factor(log(dose + 1))6.90875477931522
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.536 on 12 degrees of freedom
## Multiple R-squared:  0.5475, Adjusted R-squared:  0.359
## F-statistic: 2.904 on 5 and 12 DF,  p-value: 0.06047
```

```
anova(linmod, genmod)
```

```
## Analysis of Variance Table
##
## Model 1: colonies ~ log(dose + 1)
## Model 2: colonies ~ factor(log(dose + 1))
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     16 1881.1
## 2     12 1091.3  4    789.73 2.1709 0.1342
```

In this hypothesis test the H_0 : No lack of fit, P value is 0.1342 so we cannot reject the null no lack of fit and conclude that there is *no lack of fit*