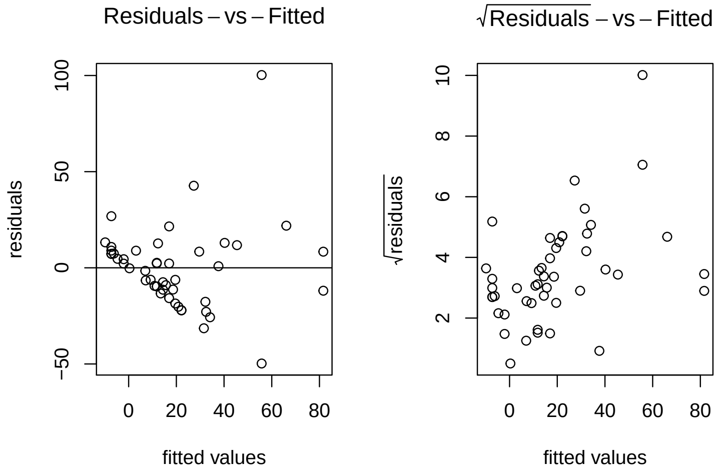
HW 4

After running a linear model with all predictors present (sex,status, income, verbal), we get insignificant values for coefficients status & verbal. As a result, an F test is run to understand whether dropping these predictors from the analysis would be beneficial. Indeed, the F test impedes us to reject "H0: Bstatus = Bverbal=0" confirming their lack of need. Hence, our final model regresses gamble only on **sex** and **income**.

```
## Analysis of Variance Table
##
## Model 1: gamble ~ sex + income
## Model 2: gamble ~ sex + status + income + verbal
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     44 22781
## 2     42 21624  2    1157.5 1.1242 0.3345
```

```
##
## Call:
## lm(formula = gamble ~ sex + income, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.757 -11.649   0.844   8.659 100.243
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.594      6.544  -2.688  0.01010 *
## sexMALE       21.634      6.809   3.177  0.00272 **
## income         5.172      0.951   5.438 2.24e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.75 on 44 degrees of freedom
## Multiple R-squared:  0.5014, Adjusted R-squared:  0.4787
## F-statistic: 22.12 on 2 and 44 DF,  p-value: 2.243e-07
```

1. One can check for constant variance by plotting residuals agains fitted values. For a better resolution one can take the square root of the abs(residuals). Clearly, constant variance assumption is not valid.



Residuals – vs – Fitted

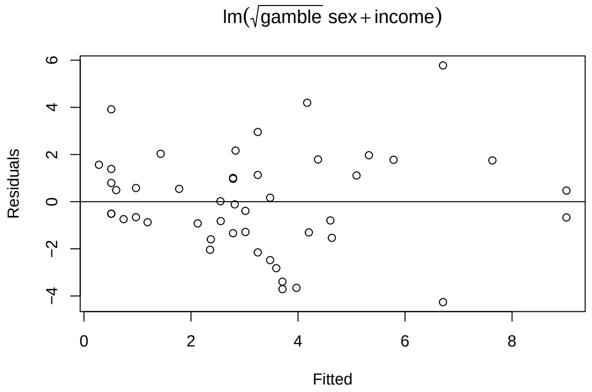
$\sqrt{\text{Residuals}}$ – vs – Fitted

Furthermore, the non constant variance can be checked by looking at the strenght of the relationship between residuals and fitted. In a constant variance scenario, the Rˆ2 should be zero.

```
lmod_res_fit= lm(sqrt(abs(residuals(lmod_siv))) ~ fitted(lmod_siv))
summary(lmod_res_fit)$r.squared
```

```
## [1] 0.1776194
```

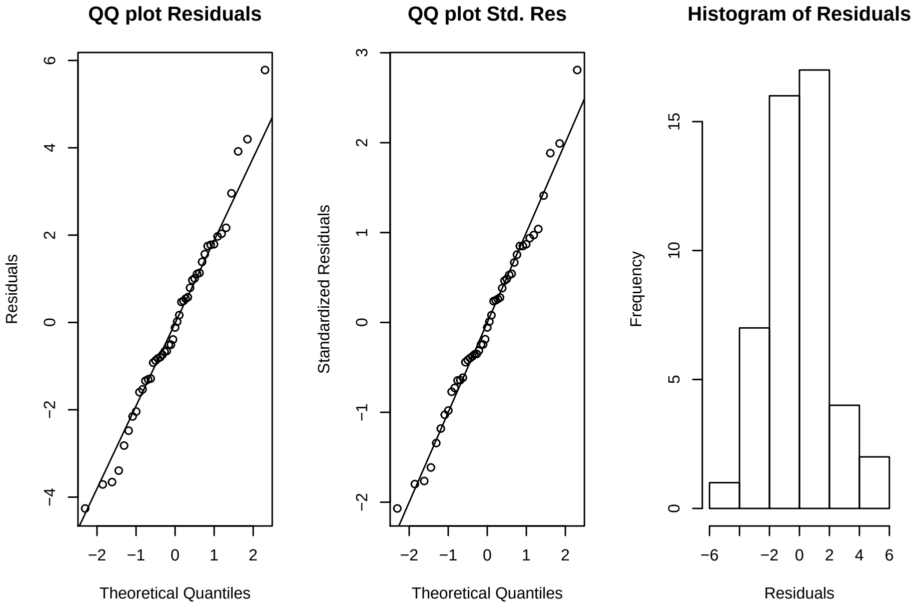
In order to “fix” this broken assumption, a transformation is applied to the response (sqrt). The constant variance does indeed check out now, as verified by the residuals vs fitted values plot.



```
## [1] "The model with a transformed response"
```

```
##
## Call:
## lm(formula = sqrt(gamble) ~ sex + income, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2597 -1.2946 -0.1159  1.2604  5.7808
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.41269    0.61708  -0.669 0.507129
## sexMALE      2.50694    0.64205   3.905 0.000321 ***
## income       0.46149    0.08968   5.146 5.95e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.146 on 44 degrees of freedom
## Multiple R-squared:  0.5164, Adjusted R-squared:  0.4945
## F-statistic:  23.5 on 2 and 44 DF,  p-value: 1.142e-07
```

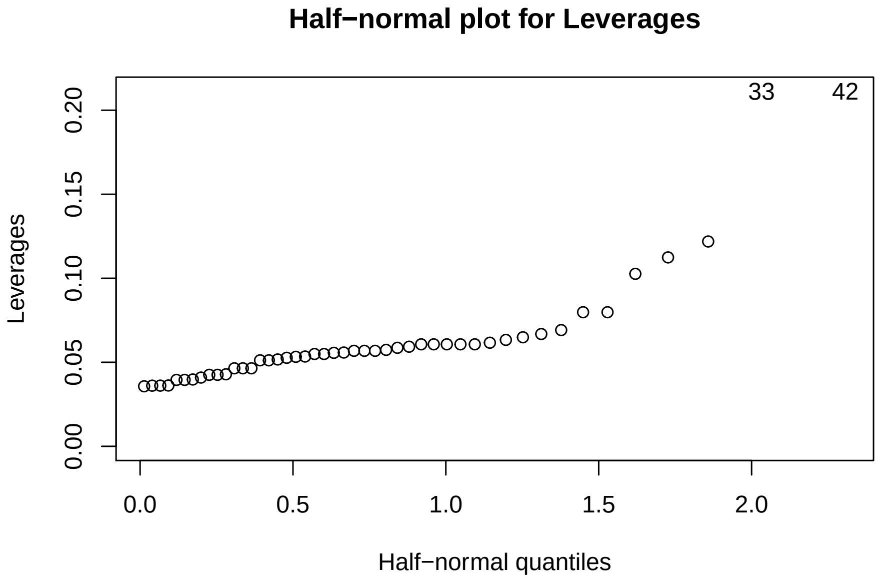
2. Normality can be checked through QQ plot (residuals/standardized), histogram , or Shapiro-Wilk (H0: residuals are normal) test. All graphical and numerical methods suggest normaility is satisfied.



```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(t_lmod_siv)
## W = 0.98407, p-value = 0.7632
```

3. Leverage points are extreme values in the X space. One can check for such points through a half plot. Any point greater than 2p/n should be investigated further. In our case observations 33 & 42 are extremes.

```
##        33        42
## 0.2112455 0.2112455
```



4. One can check for outliers by computing the studentized and apply Bonferroni's correction to find those points that do not fit the data. At an alpha level of 5% there are no apparent outliers given that maximum studentized residuals is smaller than the absolute value of the Bonferroni's critical value.

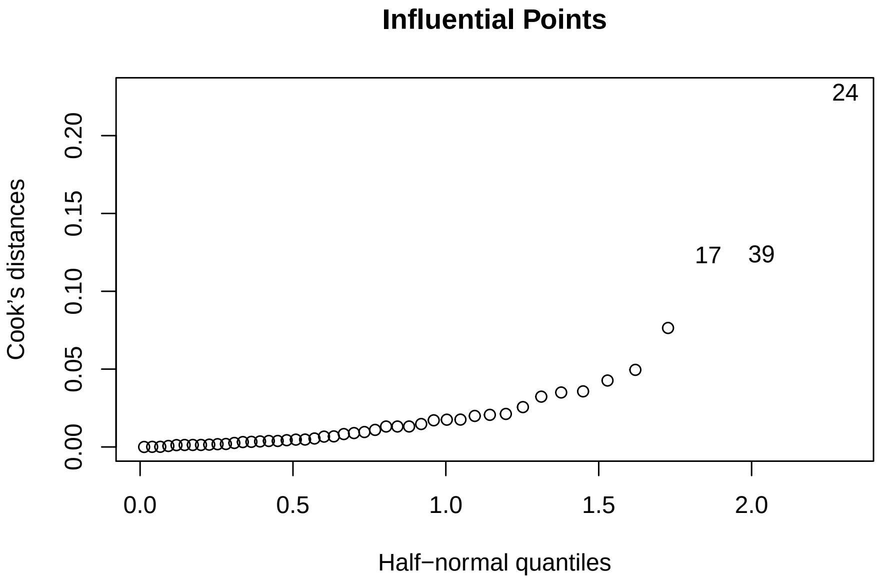
```
## [1] "Maximum Studentized Residual"
```

```
##       24
## 3.064754
```

```
#Bonferroni Critical Value (alph=0.05)
qt(.05/(47*2), t_lmod_siv$df.residual-1)
```

```
## [1] -3.510439
```

5. An influential point is a one whose removal from the data would cause a drastic change in the fit. One popular metric to identify such points is Cook's Distance.



Droppiong the "influential" points from the model does not change the coefficinets, RSE, or R^2.

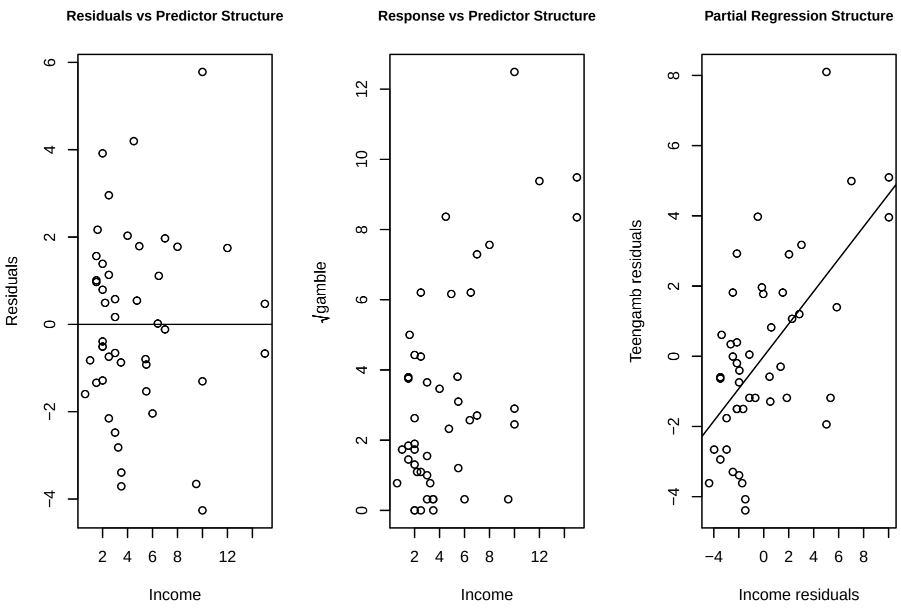
```
## lm(formula = sqrt(gamble) ~ sex + income, data = teengamb, subset = (cook <
##     max(cook)))
```

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) -0.1839095  0.5704072 -0.3224178 7.486986e-01
## sexMALE      2.3281522  0.5912681  3.9375574 2.967737e-04
## income       0.4063592  0.0841309  4.8300824 1.762443e-05
```

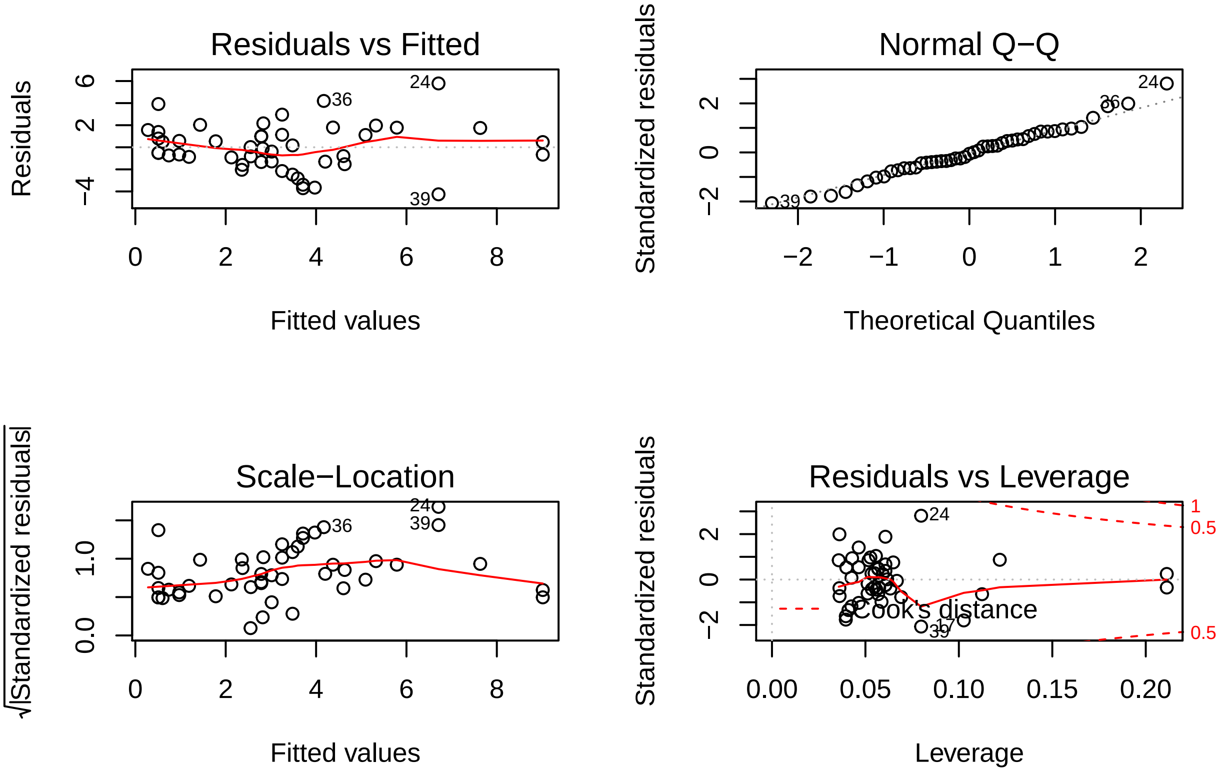
```
## [1] 1.966314
```

```
## [1] 0.4979867
```

6. There is no ambiguous (i.e. non-linear) relationship between the predictors and the response.



This is the overall structure of the model (a gift from R).


Residuals vs Fitted
Residuals
6
2
−4
0 2 4 6 8
Fitted values
24
36
39
Normal Q−Q
Standardized residuals
2
0
−2
−2 −1 0 1 2
Theoretical Quantiles
24
36
39
Scale−Location
√|Standardized residuals|
1.0
0.0
0 2 4 6 8
Fitted values
24
39
36
Residuals vs Leverage
Standardized residuals
2
0
−2
0.00 0.05 0.10 0.15 0.20
Leverage
24
39
17
Cook's distance
1
0.5
0.5