Homework 7_HW3 17_Yael Dejene Beshaw & Namit Shrivastava

```
# Firstly loading the necessary libraries
library(faraway)
```

```
Warning in check_dep_version(): ABI version mismatch:
lme4 was built with Matrix ABI version 1
Current Matrix ABI version is 0
Please re-install lme4 from source or restore original 'Matrix' package
```

```
library(ggplot2)
```

```
Warning: package 'ggplot2' was built under R version 4.3.2
```

```
library(car)
```

```
Loading required package: carData
```

```

Attaching package: 'car'
```

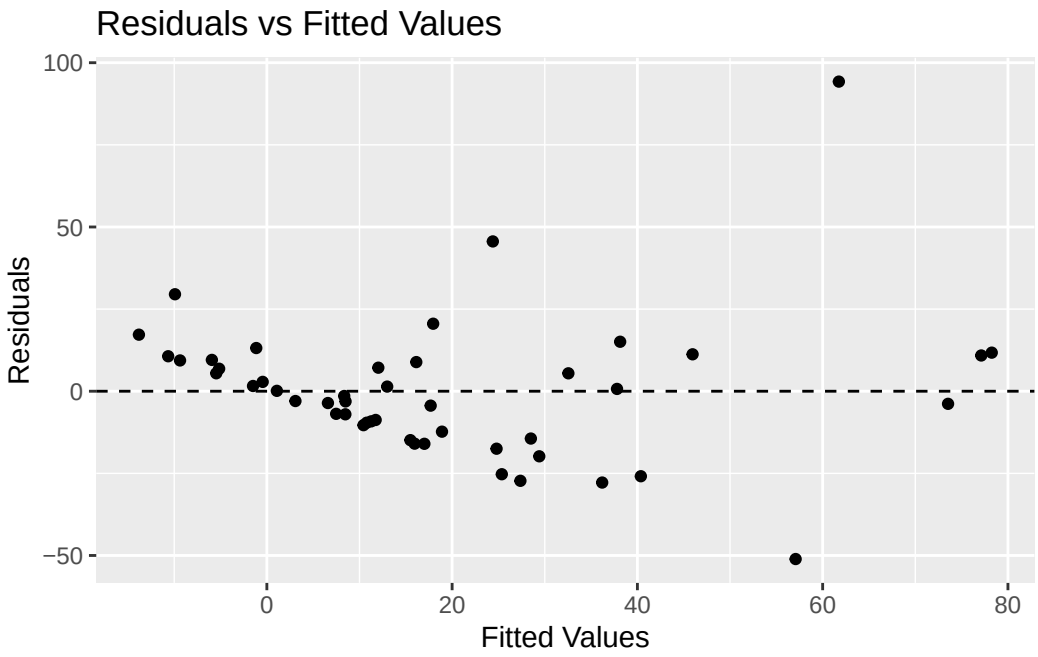
```
The following objects are masked from 'package:faraway':

    logit, vif
```

```r
# Then we load the teengamb dataset
data(teengamb)

# Fitting the linear regression model now
modelgamb <- lm(gamble ~ sex + status + income + verbal, data = teengamb)
# Summary of the model
summary(modelgamb)
```

```
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
```

1.A. Check the zero mean error assumption using residual plots

```r
residuals <- resid(modelgamb)
fitted <- fitted(modelgamb)
ggplot(data.frame(fitted, residuals), aes(x = fitted, y = residuals)) +
    geom_point() +
    geom_hline(yintercept = 0, linetype = "dashed") +
    labs(title = "Residuals vs Fitted Values", x = "Fitted Values", y = "Residuals")
```



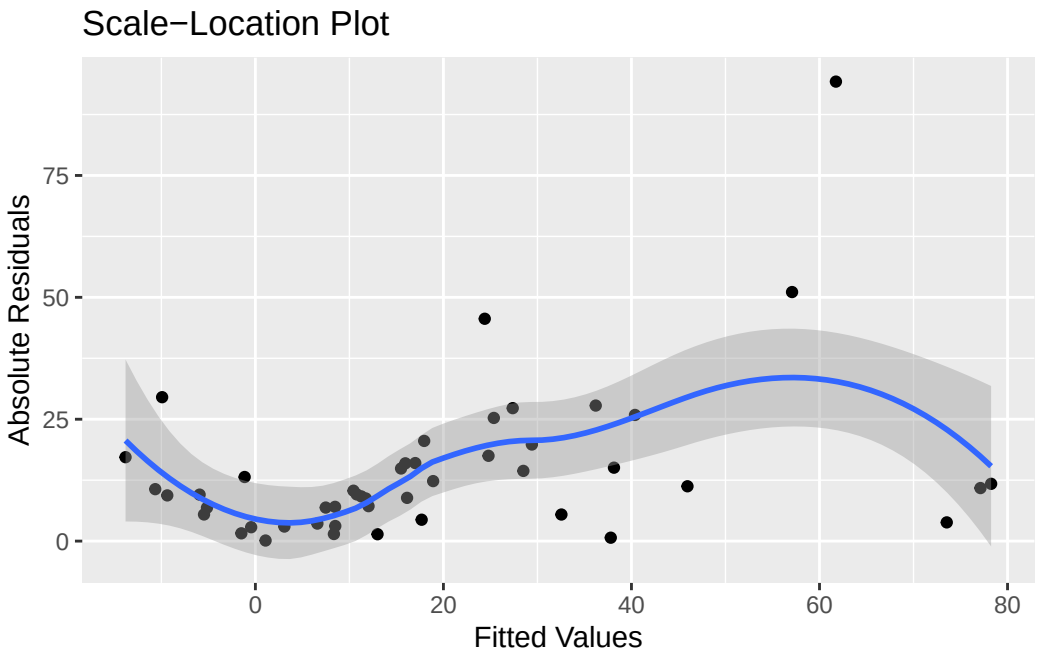
The residual plot shows the residuals (errors) on the y-axis and the fitted values on the x-axis. The zero mean error assumption is met if the residuals are randomly scattered around the horizontal line at zero.

Based on this scatter plot, we can see that there is no clear pattern like funnel or curvature, hence we can say that the model's errors have a mean of zero, and the assumption is met. Here, the model's fit is not compromised and no further suggestions are needed.

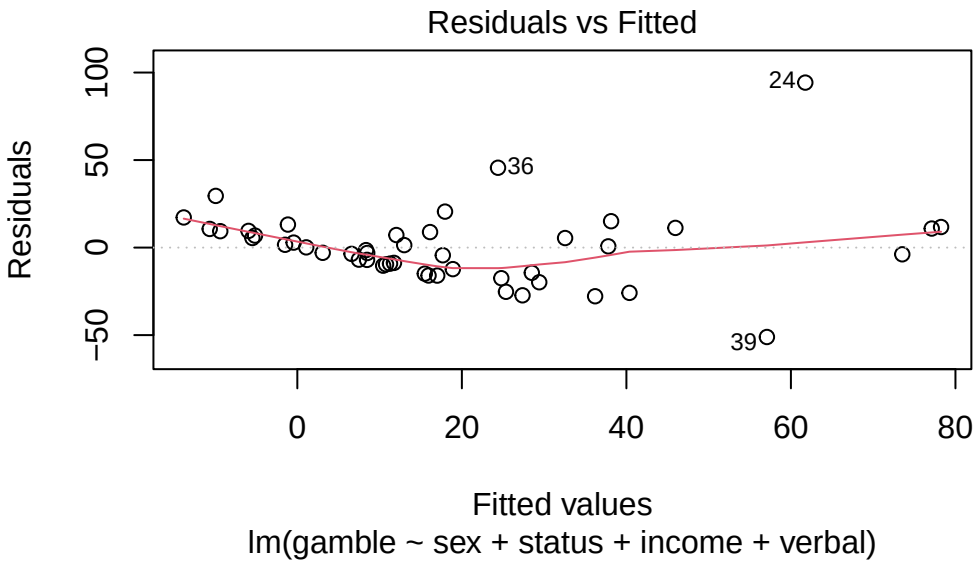
1.B. Check the constant error variance assumption using both residual plots and a formal statistical test

```
#Two ways to check the residual plots
ggplot(data.frame(fitted, residuals), aes(x = fitted, y = abs(residuals))) +
    geom_point() +
    geom_smooth(method = "loess") +
    labs(title = "Scale-Location Plot", x = "Fitted Values",
    y = "Absolute Residuals")
```

```
`geom_smooth()` using formula = 'y ~ x'
```




```
#And then using simply this

plot(modelgamb, which = 1)
```




```
# Then, performing Breusch-Pagan test
#as the formal statistical test
library(lmtest)
```

```
Loading required package: zoo

Attaching package: 'zoo'

The following objects are masked from 'package:base':

    as.Date, as.Date.numeric
```

```
bptest(modelgamb)
```

```

    studentized Breusch-Pagan test

data:  modelgamb
BP = 6.4288, df = 4, p-value = 0.1693
```

The plot is a Scale-Location plot which allows us to assess if the residuals are randomly scattered around the horizontal (zero) line. We see that there is no clear pattern in this plot and that the residuals are largly randomly scattered around zero, suggesting that the mean zero error assumption is indeed likely met.

The Breusch-Pagan test allows us to assess the constant error variance/homoscedascity assumption in which the null hypothesis is that the residuals have a constant variance across all fitted values. BP test's p value which comes out to be 0.1693, which is not significant at the 0.05 level. This confirms that we fail to reject the null hypothesis and thus there is no significant evidence of heteroscedasticity. Therefore the constant error variance assumption is likely met.

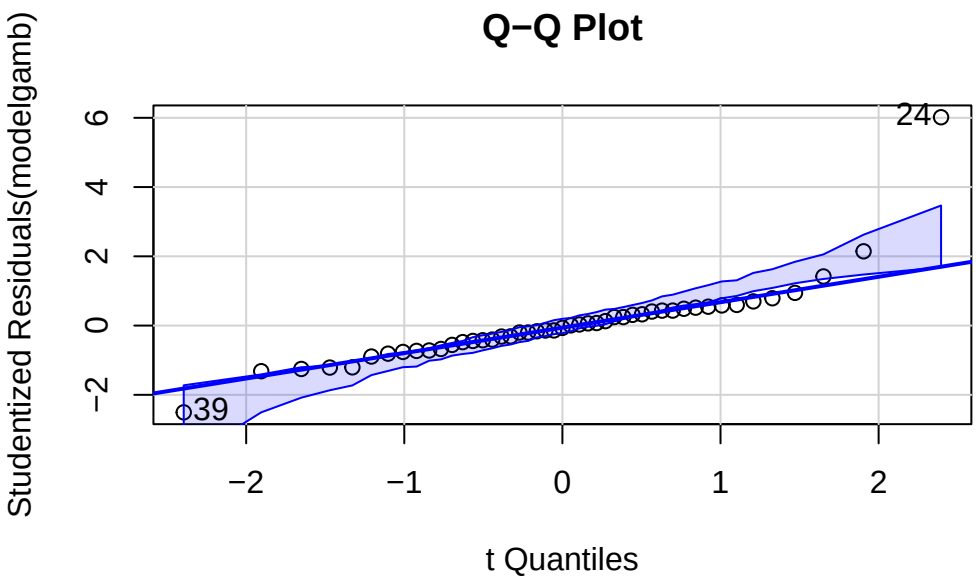
1.C. Check the error normality assumption both graphically and statistically

```
qnorm(residuals)
```

```
Warning in qnorm(residuals): NaNs produced
```

1	2	3	4	5	6	7
NaN	NaN	NaN	NaN	NaN	NaN	NaN
8	9	10	11	12	13	14
NaN	NaN	NaN	NaN	NaN	-1.1886746	NaN
15	16	17	18	19	20	21
NaN	NaN	NaN	NaN	NaN	NaN	NaN
22	23	24	25	26	27	28
NaN	NaN	NaN	0.5224919	NaN	NaN	NaN
29	30	31	32	33	34	35
NaN	NaN	NaN	NaN	NaN	NaN	NaN
36	37	38	39	40	41	42
NaN	NaN	NaN	NaN	NaN	NaN	NaN
43	44	45	46	47		
NaN	NaN	NaN	NaN	NaN		

```
qqPlot(modelgamb, main = "Q-Q Plot")
```




```
[1] 24 39
```

We observe that the points in the Q-Q plot roughly follow the diagonal line, this suggests that the residuals are normally distributed.

The plot also highlights that point 24 and 39 are seen to be potential outliers which may affect the normality of the residuals. We must conduct further investigations on these points in order to assess their influence on the model.

Perform Shapiro-Wilk test

```
shapiro.test(residuals)
```

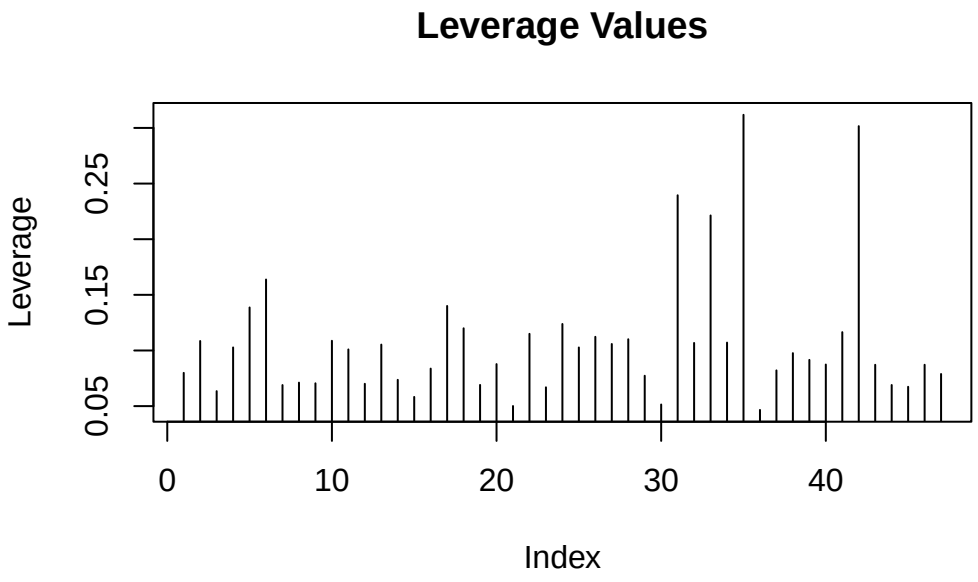
```
    Shapiro-Wilk normality test

data:  residuals
W = 0.86839, p-value = 8.16e-05
```

For this test we have the null hypothesis as; residuals are normally distributed and alternate hypothesis as; the residuals are not normally distributed.

Based on the result obtained in the Shapiro-Wilk test, we see that the p-value = 8.16e-05 is less than 0.05, and so we reject the null hypothesis. We can thus state that there is statistically significant evidence of deviation from normality.

1.D. Check for observations with large leverage

```
leverage <- hatvalues(modelgamb)
plot(leverage, type = "h", main = "Leverage Values", ylab = "Leverage")
```



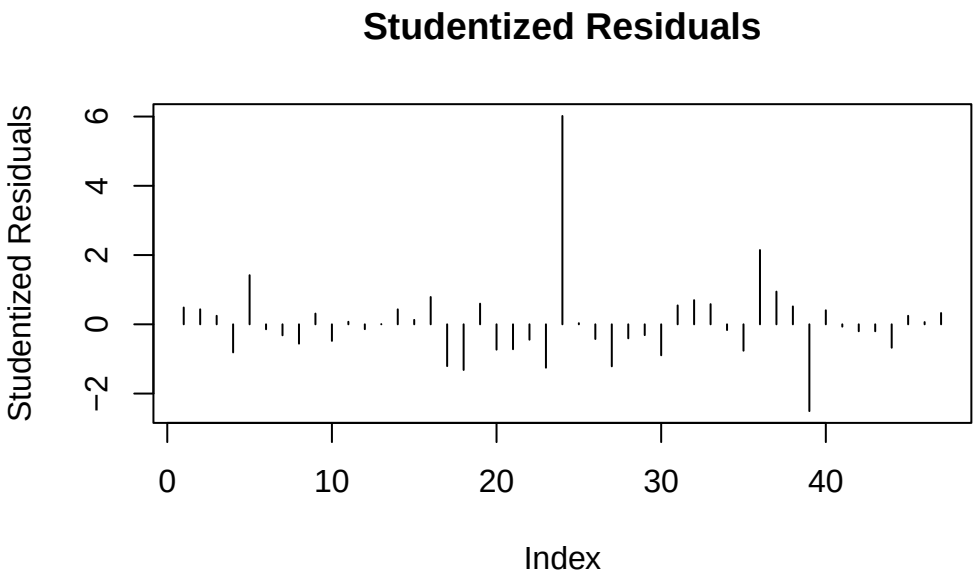
As discussed in class, we know that the leverage values measure the influence of each observation on the fitted values. High leverage indicates to us that there is disproportionate influence of certain observations.

Based on the plot, the observations such as 35 and 42 have high leverage and should be investigated further to understand their influence on the model. Investigations on these points would allow us to assess if they are outliers or data errors that may need to be removed from the model.

By identifying these high leverage points, we can actually ensure that the regression model is not unduly influenced by a few observations and improve the model's reliability and robustness.

1.E. Check for outliers

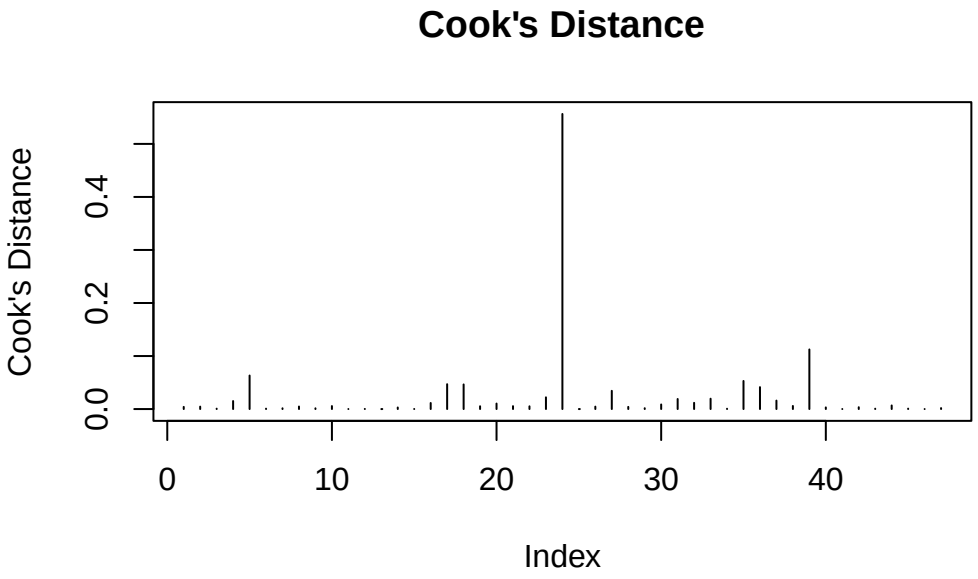
```
studentized_residuals <- rstudent(modelgamb)
plot(studentized_residuals, type = "h", main = "Studentized Residuals", ylab = "Studentize
```



Seeing the general trend, we can see that the outliers seem to be those observations which have studentized residuals beyond +/-2. There are only two of these indices and further investigations need to be done to understand their influence on the model. Similar to the observations with large leverage, the residuals beyond +/-2 may indicate to us that there may be errors in the data or that the model does not fit well at these specified indices.

1.F. Check for influential points

```
cooksd <- cooks.distance(modelgamb)
plot(cooksd, type = "h", main = "Cook's Distance", ylab = "Cook's Distance")
```

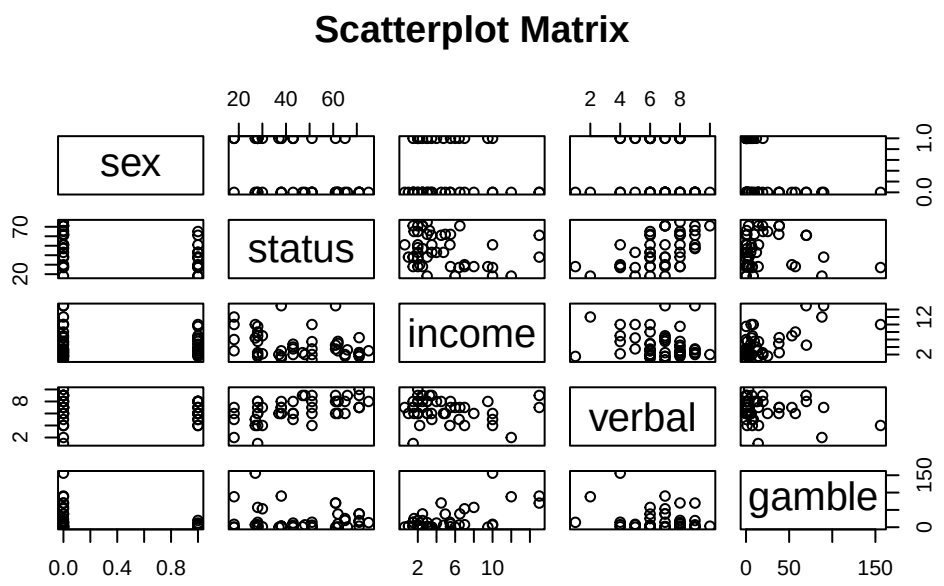


Based on the result, we see that index point 24 has the greatest Cook's Distance relative to the other indices. Cook's Distance allows us to identifying influential points and thus ensure that the regression model is unduly influenced by a few observations. The index point 24 indicates that the observation is influential on the model and it's deletion would impact the model more than the deletion of other data points.

In the previous questions, we have seen that observations such as index points 35 and 42 seem to be outliers that could potentially impact our data, in this plot, we see that these points do not have as much negative influence on the model as we suspect and thus we would be able to continue with these points. However, this results leads us to conduct further investigations on index point 24 to see if we may need to remove it as it was also an outlier in our Q-Q residual plot, potentially impacting the model's robustness, reliability, and fit.

1.G. Check the structure of relationship between the predictors and the response

```
pairs(teengamb, main = "Scatterplot Matrix")
```



Upon examining the structure of relationship between the predictors and response, we see that there are not many clear signs of a relationship between the predictors and response. Gamble vs Income has the clearest trend where we see a positive association indicating to us that income is an important predictor for gamble, which corroborated in our initial regression analysis where the coefficient for income is 4.962 with a statistically significant p-value of 1.79e-05.