MATH3424 HW3

Chow, Hau Cheung Jasper (hcjchow / 20589533)

October 25, 2021

Q1

Q1a) Number of df's in ANOVA F-test $= 48 = n - p - 1$ where $p = 1$ in this case, hence number of workers is $n = 50$.

Q1b) $Var(Y) = \sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n-1} = SST/(n-1) = (SSE+SSR)/(n-1) = (98.8313+338.449)/49 = 8.924088$.

Q1c) Since it is well known $\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}$ in simple linear regression, $\bar{Y} = 15.58 + 0.52 * (-2.81) = 14.1188$

Q1d) Since $n = 50$, if $\bar{X} = 0.52$ then $\sum_{i=1}^{50} X_i = 0.52 * 50 = 26$

Hence there are 26 men and 24 women.

Q1e) Percentage of variability between Y and X can be expressed as $R^2 = SSR/SST = 98.8313/(98.8313 + 338.449) = 0.2260136$ so the answer is 22.60136%.

Q1f) $[Cor(Y, X)]^2 = R^2$

Since the regression coefficient for X is negative, we choose the negative root. $Cor(Y, X) = -(0.2261036)^{1/2} = -0.4754089$

Q1g) $\hat{\beta}_1$ describes: estimated weekly wages of a male worker minus estimated weekly wages of a female worker (the male worker earns \$281 less).

Q1h) $15.58 - 2.81 = 12.77$. Then multiply by \$100 to get \$1277

Q1i) \$1558

Q1j) CI is $\hat{\beta}_1 \pm t_{n-2,\alpha/2} \times s.e.(\hat{\beta}_1)$

$t_{n-2,\alpha/2} = t_{48,0.025} = 2.010635, s.e.(\hat{\beta}_1) = 0.75$

$CI = [-2.81 \pm 2.010635 * 0.75] = [-4.317976, -1.302024]$

Q1k) $H_0 : \hat{\beta}_1 = 0, H_1 : \hat{\beta}_1 \neq 0$

$$t_1 = \frac{\hat{\beta}_1}{s.e.(\hat{\beta}_1)} = \frac{-2.81}{0.75} = -3.7467$$

$t_{critical} = t_{n-p-1,\alpha/2} = t_{48,0.025} = 2.010635$

Since $|t_1| > t_{critical}$, we reject H_0 at significance level $\alpha = 0.05$ i.e. we conclude that the average weekly ages of men is not equal to that of the women.

```r
# Q1b
(98.8313+338.449)/49
```

```
## [1] 8.924088
```

```r
# Q1c
15.58+0.52*(-2.81)
```

```
## [1] 14.1188
```

```
# Q1e, Q1f, Q1h
98.8313/(98.8313+338.449)
```

```
## [1] 0.2260136
```

```
(98.8313/(98.8313+338.449))^0.5
```

```
## [1] 0.4754089
```

```
15.58-2.81
```

```
## [1] 12.77
```

```
# Q1j
q1j_t = qt(0.025,df=48,lower.tail=FALSE)
q1j_t
```

```
## [1] 2.010635
```

```
-2.81-q1j_t*0.75
```

```
## [1] -4.317976
```

```
-2.81+q1j_t*0.75
```

```
## [1] -1.302024
```

Q2

```
setwd("/Users/jchow/Downloads/MATH3424 R") # need to set the starting directory as this
elect_data <- read.table(file="Presidential Election Data.txt",header=TRUE)

elect_data$GI <- elect_data$G * elect_data$I
elect_data_2a <- subset(elect_data,select=c("V","I","D","W","GI","P","N"))

elect_model_1 = lm(V ~ ., data = elect_data_2a)
summary(elect_model_1)
```

```
##
## Call:
## lm(formula = V ~ ., data = elect_data_2a)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.041742 -0.021066 -0.003611  0.011760  0.087914
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.5111627  0.0321992  15.875 2.40e-10 ***
## I           -0.0201077  0.0168979  -1.190   0.2539
## D            0.0546159  0.0205705   2.655   0.0188 *
## W            0.0133905  0.0422639   0.317   0.7560
## GI           0.0096901  0.0017712   5.471 8.24e-05 ***
## P           -0.0007224  0.0040046  -0.180   0.8594
## N           -0.0051822  0.0038083  -1.361   0.1951
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.04113 on 14 degrees of freedom
## Multiple R-squared:  0.7898, Adjusted R-squared:  0.6998
## F-statistic: 8.769 on 6 and 14 DF,  p-value: 0.0004347
```

```
cor(elect_data_2a$I, elect_data_2a$D)
```

```
## [1] 0.8174431
```

Q2a) At $D = 1:$

$\hat{V} = 0.5658 - 0.0201I + 0.0134W + 0.0097(G \cdot I) - 0.0007P - 0.0052N$

At $D = 0:$

$\hat{V} = 0.5112 - 0.0201I + 0.0134W + 0.0097(G \cdot I) - 0.0007P - 0.0052N$

At $D = -1:$

$\hat{V} = 0.4566 - 0.0201I + 0.0134W + 0.0097(G \cdot I) - 0.0007P - 0.0052N$

The coefficient for D represents the increase in predicted democratic share of the vote when:

a non-Democrat and non-Republican incumbent running for election (0) compared to if a Republican incumbent running for election (-1) AND all other variables held constant; OR

a Democrat incumbent running for election (1) compared to if a non-Democrat and non-Republican incumbent running for election (0) AND all other variables held constant.

Q2b) Only the coefficients for D and $G \cdot I$ are statistically significant at $\alpha = 0.05$ due to their low p-values. As such, we probably don't need to keep I. Furthermore, I and D have high correlation of 0.81 so we may not need to include I if D already in the model.

Q2c) We should keep it, since the p-value for the coefficient of $G \cdot I$ is very low, making it statistically significant at $\alpha = 0.05$. However due to high correlation of D and I as mentioned in 2b, it may be the case that we should include $G \cdot D$ instead of $G \cdot I$.

Q2d) As stated in 2b, we may not need to include both I and D since they are highly correlated. Here we choose to include D and exclude I since its correlation with V is higher. Secondly, the indicator variable W indicating presence of election preceding or following war may not be of much use since only 3 observations have $W = 1$, hence we may consider dropping it. Additionally, if a given party is incumbent and performs well while in office (ie good economic performance aka high values of N and P), we expect that party to remain incumbent (have a higher vote share.) Therefore, we expect that the interaction terms $P \cdot D$ and $N \cdot D$ to be significant. Fourth, we have the interaction term $G \cdot D$ replacing $G \cdot I$ and the term G itself.

Hence, we remove I, remove W, add G, replace $G \cdot I$ with $G \cdot D$ and add interaction terms $P \cdot D$ and $N \cdot D$. This leaves us with the model $V = \beta_0 + \beta_1 D + \beta_2(G \cdot D) + \beta_3 P + \beta_4 N + \beta_5(P \cdot D) + \beta_6(N \cdot D) + \beta_7 G + \epsilon$.

```
cor(elect_data_2a$V, elect_data_2a$D)
```

```
## [1] 0.49917
```

```
cor(elect_data_2a$V, elect_data_2a$I)
```

```
## [1] 0.3464773
```

```
elect_data$GD <- elect_data$G * elect_data$D
elect_data$PD <- elect_data$P * elect_data$D
elect_data$ND <- elect_data$N * elect_data$D
elect_data_2d <- subset(elect_data,select=c("V","D","GD","P","N","PD","ND","G"))

elect_model_2 = lm(V ~ ., data = elect_data_2d)
summary(elect_model_2)
```

```
##
## Call:
## lm(formula = V ~ ., data = elect_data_2d)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.064490 -0.023202  0.003524  0.024861  0.041884
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.525094   0.026532  19.791 4.34e-11 ***
## D           -0.007028   0.031802  -0.221 0.828537
## GD           0.010224   0.002386   4.285 0.000887 ***
## P           -0.002191   0.003215  -0.682 0.507493
## N           -0.008774   0.003435  -2.554 0.023989 *
## PD          -0.002299   0.005010  -0.459 0.653957
## ND           0.008089   0.003465   2.335 0.036243 *
## G            0.002815   0.002272   1.239 0.237245
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03852 on 13 degrees of freedom
## Multiple R-squared:  0.8288, Adjusted R-squared:  0.7367
## F-statistic: 8.993 on 7 and 13 DF,  p-value: 0.0004067
```

As we can see, some coefficients are not statistically significant, like PD, G, D, and P. If we remove the predictors one at a time starting with the one whose coefficient has the highest p-value (D) and refit the model:

```
elect_data_2d <- subset(elect_data,select=c("V","GD","P","N","PD","ND","G"))
elect_model_3 = lm(V ~ ., data = elect_data_2d)
summary(elect_model_3)
```

```
##
## Call:
## lm(formula = V ~ ., data = elect_data_2d)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.064272 -0.023259  0.004085  0.025675  0.040699
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.524489   0.025478  20.586 7.27e-12 ***
## GD           0.010074   0.002209   4.561 0.000444 ***
## P           -0.002091   0.003073  -0.680 0.507311
## N           -0.008698   0.003299  -2.636 0.019547 *
## PD          -0.003088   0.003394  -0.910 0.378325
## ND           0.007668   0.002796   2.743 0.015865 *
## G            0.002780   0.002188   1.271 0.224596
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03719 on 14 degrees of freedom
## Multiple R-squared:  0.8282, Adjusted R-squared:  0.7546
## F-statistic: 11.25 on 6 and 14 DF,  p-value: 0.0001149
```

Keep going until all coefficients are significant. Remove P next.

```
elect_data_2d <- subset(elect_data,select=c("V","GD","N","PD","ND","G"))
elect_model_4 = lm(V ~ ., data = elect_data_2d)
summary(elect_model_4)
```

```
## 
## Call:
## lm(formula = V ~ ., data = elect_data_2d)
## 
## Residuals:
##       Min        1Q    Median        3Q       Max 
## -0.066030 -0.017380  0.003532  0.023745  0.043756 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  0.514047   0.019971  25.740 7.94e-14 ***
## GD           0.010493   0.002083   5.037 0.000147 ***
## N           -0.008833   0.003234  -2.731 0.015456 *  
## PD          -0.003581   0.003256  -1.100 0.288740    
## ND           0.007656   0.002745   2.789 0.013764 *  
## G            0.003721   0.001666   2.233 0.041203 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.03652 on 15 degrees of freedom
## Multiple R-squared:  0.8225, Adjusted R-squared:  0.7634 
## F-statistic:  13.9 on 5 and 15 DF,  p-value: 3.522e-05
```

Remove PD next.

```
elect_data_2d <- subset(elect_data,select=c("V","GD","N","ND","G"))
elect_model_5 = lm(V ~ ., data = elect_data_2d)
summary(elect_model_5)
```

```
## 
## Call:
## lm(formula = V ~ ., data = elect_data_2d)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.06132 -0.01965  0.00500  0.02325  0.05423 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  0.510947   0.019900  25.676 1.97e-14 ***
## GD           0.011379   0.001933   5.885 2.30e-05 ***
## N           -0.008288   0.003217  -2.577   0.0203 *  
## ND           0.005288   0.001714   3.085   0.0071 ** 
## G            0.004099   0.001641   2.498   0.0238 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.03675 on 16 degrees of freedom
## Multiple R-squared:  0.8082, Adjusted R-squared:  0.7603
## F-statistic: 16.86 on 4 and 16 DF,  p-value: 1.367e-05
```

```
anova(elect_model_5, elect_model_2)
```

```
## Analysis of Variance Table
##
## Model 1: V ~ GD + N + ND + G
## Model 2: V ~ D + GD + P + N + PD + ND + G
##   Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1     16 0.021615
## 2     13 0.019289  3 0.0023258 0.5225 0.6743
```

Ultimately, this leaves us with the model $\hat{V} = 0.5109 + 0.0114(G \cdot D) - 0.0083N + 0.0053(N \cdot D) + 0.0041G$ as our final model, in which each coefficient is statistically significant at the $\alpha = 0.05$ level, and the adjusted R-squared is reasonably high at 0.76 and is a slight increase over the original full model $V = \beta_0 + \beta_1 D + \beta_2(G \cdot D) + \beta_3 P + \beta_4 N + \beta_5(P \cdot D) + \beta_6(N \cdot D) + \beta_7 G + \epsilon$.

Q3

```
elect_data_q3 <- read.table(file="Presidential Election Data.txt",header=TRUE)

elect_data_q3$GI <- elect_data_q3$G * elect_data_q3$I

elect_data_q3$D <- as.factor(elect_data_q3$D)  # D in {-1, 0, 1} is indicator
elect_data_q3 <- within(elect_data_q3, D<-relevel(D,ref=2)) # 0 aka 2nd category is the base class
elect_data_q3a <- subset(elect_data_q3,select=c("V","I","D","W","GI","P","N"))

elect_model_q3 = lm(V ~ ., data = elect_data_q3a)
summary(elect_model_q3)
```

```
##
## Call:
## lm(formula = V ~ ., data = elect_data_q3a)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.044201 -0.022728 -0.002548  0.011671  0.084681
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.5054760  0.0364190  13.879 3.58e-09 ***
## I           -0.0205982  0.0174858  -1.178 0.259912
## D-1         -0.0469714  0.0291912  -1.609 0.131600
## D1           0.0633485  0.0312177   2.029 0.063423 .
## W            0.0123948  0.0436938   0.284 0.781127
## GI           0.0094222  0.0019580   4.812 0.000339 ***
## P           -0.0006963  0.0041333  -0.168 0.868808
## N           -0.0051083  0.0039349  -1.298 0.216773
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04245 on 13 degrees of freedom
## Multiple R-squared:  0.7922, Adjusted R-squared:  0.6802
## F-statistic: 7.078 on 7 and 13 DF,  p-value: 0.001307
```

Q3a) At $D = 1$:

$\hat{V} = 0.5688 - 0.0206I + 0.0124W + 0.0094(G \cdot I) - 0.0007P - 0.0051N$

At $D = 0$:

$\hat{V} = 0.5055 - 0.0206I + 0.0124W + 0.0094(G \cdot I) - 0.0007P - 0.0051N$

At $D = -1$:

$\hat{V} = 0.4585 - 0.0206I + 0.0124W + 0.0094(G \cdot I) - 0.0007P - 0.0051N$

α_1, regression coefficient of D_1, corresponds to the increase in democratic vote share if Democrat incumbent running for election (1) compared to if a non-Democrat and non-Republican incumbent running for election (0) AND all other variables held constant.

α_2, regression coefficient of D_2, corresponds to the decrease in democratic vote share if Republican incumbent running for election (-1) compared to if non-Democrat and non-Republican incumbent running for election (0), AND all other variables held constant.

Q3b) Consider the initial model $V = \beta_0 + \beta_1 I + \beta_2 D + \beta_3 W + \beta_4 (G \cdot I) + \beta_5 P + \beta_6 N + \epsilon$. Since $D \in \{-1, 0, 1\}$, consider the initial model for all values of D:

$V = \beta_0 + \beta_1 I + \beta_2 + \beta_3 W + \beta_4 (G \cdot I) + \beta_5 P + \beta_6 N + \epsilon$ at $D = 1$

$V = \beta_0 + \beta_1 I + 0 + \beta_3 W + \beta_4 (G \cdot I) + \beta_5 P + \beta_6 N + \epsilon$ at $D = 0$

$V = \beta_0 + \beta_1 I - \beta_2 + \beta_3 W + \beta_4 (G \cdot I) + \beta_5 P + \beta_6 N + \epsilon$ at $D = -1$

Consider our new model $V = \beta_0 + \beta_1 I + \alpha_1 D_1 + \alpha_2 D_2 + \beta_3 W + \beta_4 (G \cdot I) + \beta_5 P + \beta_6 N + \epsilon$. Now assume in the new model $\alpha_1 = -\alpha_2$. Then new model becomes $V = \beta_0 + \beta_1 I - \alpha_2 D_1 + \alpha_2 D_2 + \beta_3 W + \beta_4 (G \cdot I) + \beta_5 P + \beta_6 N + \epsilon = \beta_0 + \beta_1 I + \alpha_2 (D_2 - D_1) + \beta_3 W + \beta_4 (G \cdot I) + \beta_5 P + \beta_6 N + \epsilon$. Now consider the new model for all values of D.

$V = \beta_0 + \beta_1 I - \alpha_2 + \beta_3 W + \beta_4 (G \cdot I) + \beta_5 P + \beta_6 N + \epsilon$ at $D = 1$ since $D_2 - D_1 = -1$;

$V = \beta_0 + \beta_1 I + 0 + \beta_3 W + \beta_4 (G \cdot I) + \beta_5 P + \beta_6 N + \epsilon$ at $D = 0$ since $D_2 - D_1 = 0$;

$V = \beta_0 + \beta_1 I + \alpha_2 + \beta_3 W + \beta_4 (G \cdot I) + \beta_5 P + \beta_6 N + \epsilon$ at $D = -1$ since $D_2 - D_1 = 1$;

Observe that if $\alpha_2 = -\beta_2$, we simply get the equation(s) for the initial model in each case.

Q3c) Let $V = \beta_0 + \beta_1 I + \alpha_1 D_1 + \alpha_2 D_2 + \beta_3 W + \beta_4 (G \cdot I) + \beta_5 P + \beta_6 N + \epsilon$ be the new model. Declare $\gamma = \alpha_1 + \alpha_2$. Then the new model can be rewritten as $V = \beta_0 + \beta_1 I + (\gamma - \alpha_2) D_1 + \alpha_2 D_2 + \beta_3 W + \beta_4 (G \cdot I) + \beta_5 P + \beta_6 N + \epsilon = \beta_0 + \beta_1 I + \gamma * D_1 + \alpha_2 * (D_2 - D_1) + \beta_3 W + \beta_4 (G \cdot I) + \beta_5 P + \beta_6 N + \epsilon$. As such, the null hypothesis $H_0 : \alpha_1 = -\alpha_2$ can be rewritten as: $H_0 : \gamma = 0$ with $H_1 : \gamma \neq 0$ as alternative.

```r
elect_data_3c <- subset(elect_data,select=c("V","I","D","W","GI","P","N"))

elect_data_3c$D2 <- rep(0,dim(elect_data)[1])
elect_data_3c$D2[which(elect_data_3c$D==-1)] <- 1 # create D2 column

elect_data_3c$D1 <- rep(0,dim(elect_data)[1])
elect_data_3c$D1[which(elect_data_3c$D==1)] <- 1 # create D1 column
elect_data_3c$D2minusD1 <- elect_data_3c$D2 - elect_data_3c$D1 # create D2-D1 column

elect_data_3c <- subset(elect_data_3c,select=c("V","I","D1","D2minusD1","W","GI","P","N"))
elect_model_3c = lm(V ~ ., data = elect_data_3c)
summary(elect_model_3c)
```

```
##
## Call:
## lm(formula = V ~ ., data = elect_data_3c)
##
```

```
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.044201 -0.022728 -0.002548  0.011671  0.084681
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.5054760  0.0364190  13.879 3.58e-09 ***
## I           -0.0205982  0.0174858  -1.178 0.259912
## D1           0.0163771  0.0429257   0.382 0.708979
## D2minusD1   -0.0469714  0.0291912  -1.609 0.131600
## W            0.0123948  0.0436938   0.284 0.781127
## GI           0.0094222  0.0019580   4.812 0.000339 ***
## P           -0.0006963  0.0041333  -0.168 0.868808
## N           -0.0051083  0.0039349  -1.298 0.216773
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04245 on 13 degrees of freedom
## Multiple R-squared:  0.7922, Adjusted R-squared:  0.6802
## F-statistic: 7.078 on 7 and 13 DF,  p-value: 0.001307
```

From the t-test on coefficient of D_1 aka γ, the p-value is very large so we fail to reject H_0 at all significance below 0.708979. Hence the data does seem to support the assumption that $\alpha_1 = -\alpha_2$.

Q4

Q4a) From the table, correlation coefficient i.e. $Cor(X, Y) = Cor(X, Y^\lambda) = -0.777$ when $\lambda = 1$.

Q4b) The most important objective of variable transformation is to achieve linearity i.e. a high magnitude of correlation coefficient. When $\lambda > 0$, the correlation is highly negative. When $\lambda < 0$, the correlation is highly positive. The best value for λ is -1, since the absolute value of the correlation is the highest among all tested λ. This implies that X and Y^{-1} are extremely strongly linearly related.

Q4c) $\frac{1}{Y} = \beta_0 + \beta_1 X + \epsilon$

Q5

The standard regression assumptions are:

(1) Assumption of linearity

(2.1) Assumption that errors are independent of each other

(2.2) Assumption that errors are normally distributed

(2.3) Assumption that errors each have mean 0

(2.4) Assumption that errors each have common variance σ^2

(3.1) Assumption that predictor variables $X_1, X_2, ...X_n$ are nonrandom.

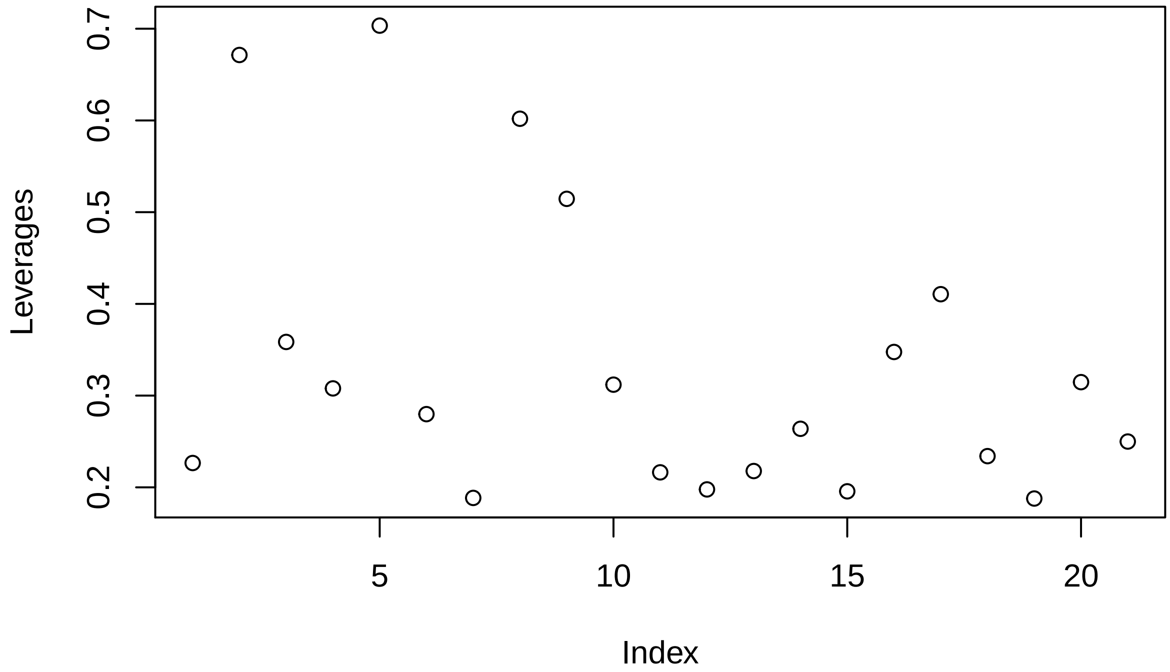
(3.2) Assumption that predictor values $x_{1j}, x_{2j}, ..., x_{nj}$ are measured without error.

(3.3) Assumption that predictors $X_1, X_2, ...X_n$ are independent of each other.

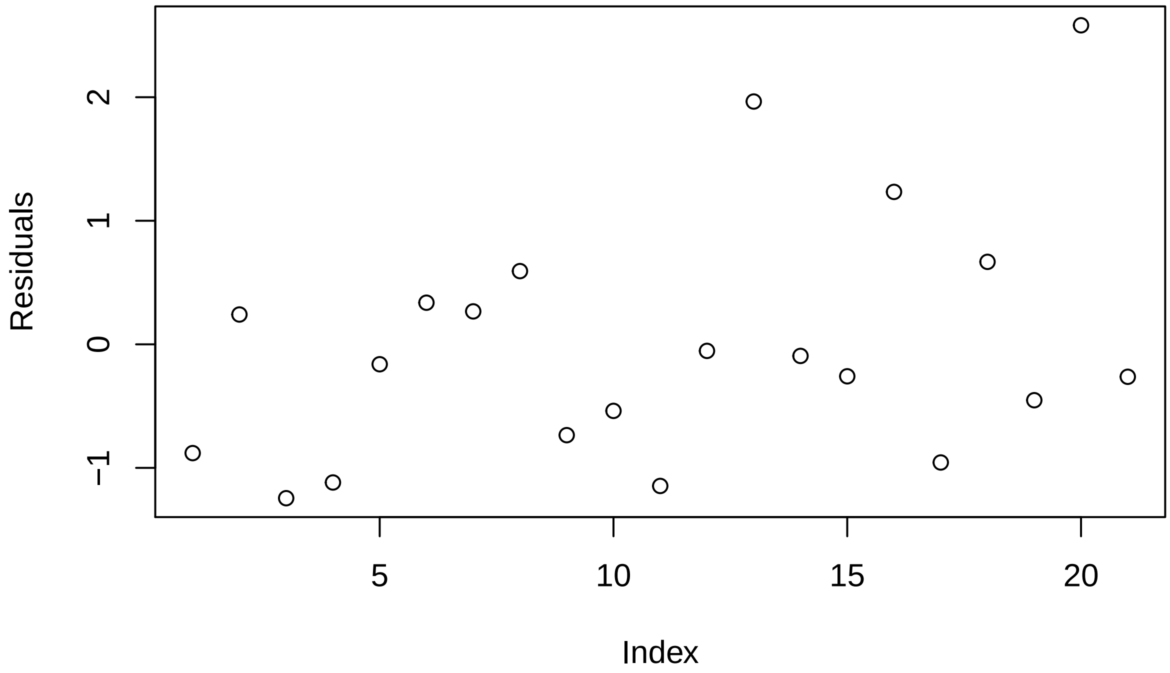
(4) Assumption that all observations are equally reliable and have an approximately equal role in determining regression results.

```r
leverages <- hatvalues(elect_model_1) # compute leverages for model V
elect_model_1.residuals <- rstandard(elect_model_1)
```

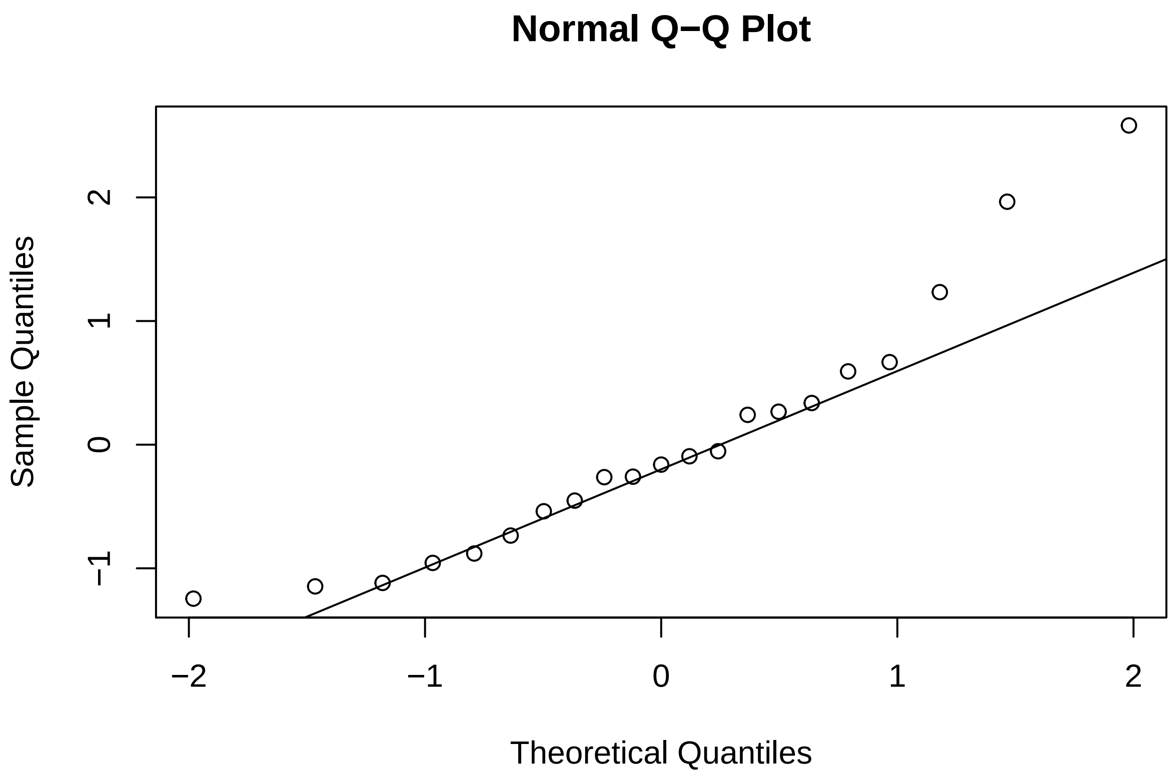
```r
plot(seq(1,length(elect_data$V)), leverages, xlab="Index", ylab="Leverages")
```



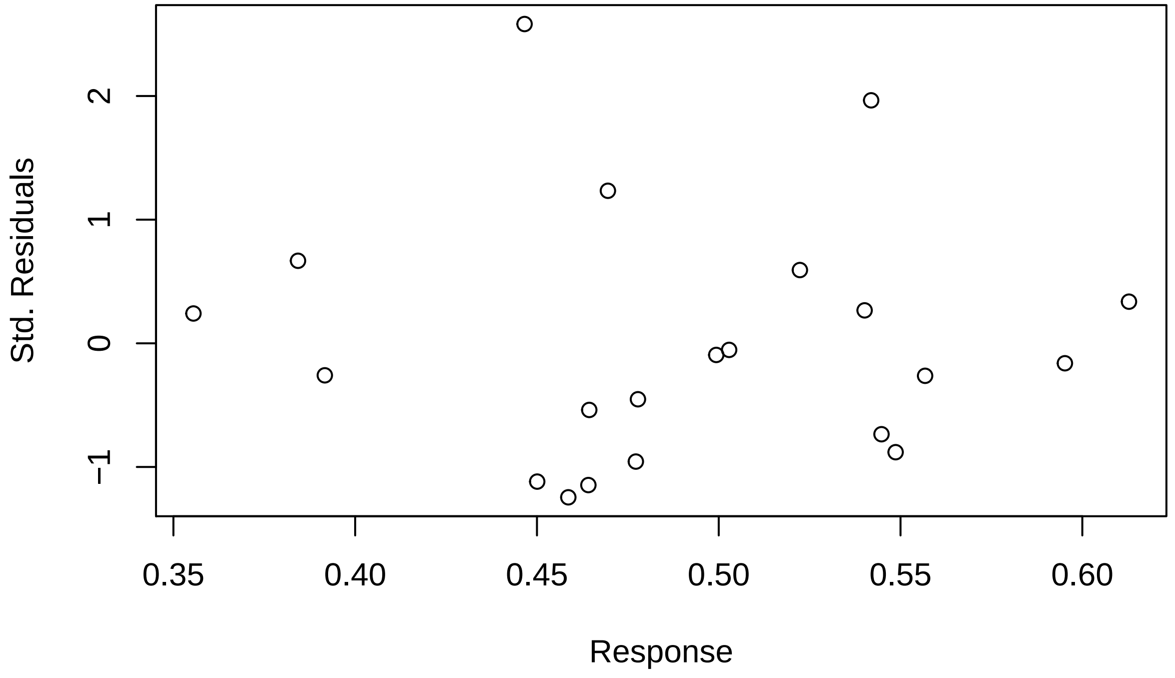
```r
plot(seq(1,length(elect_data$V)), elect_model_1.residuals, xlab="Index", ylab="Residuals")
```



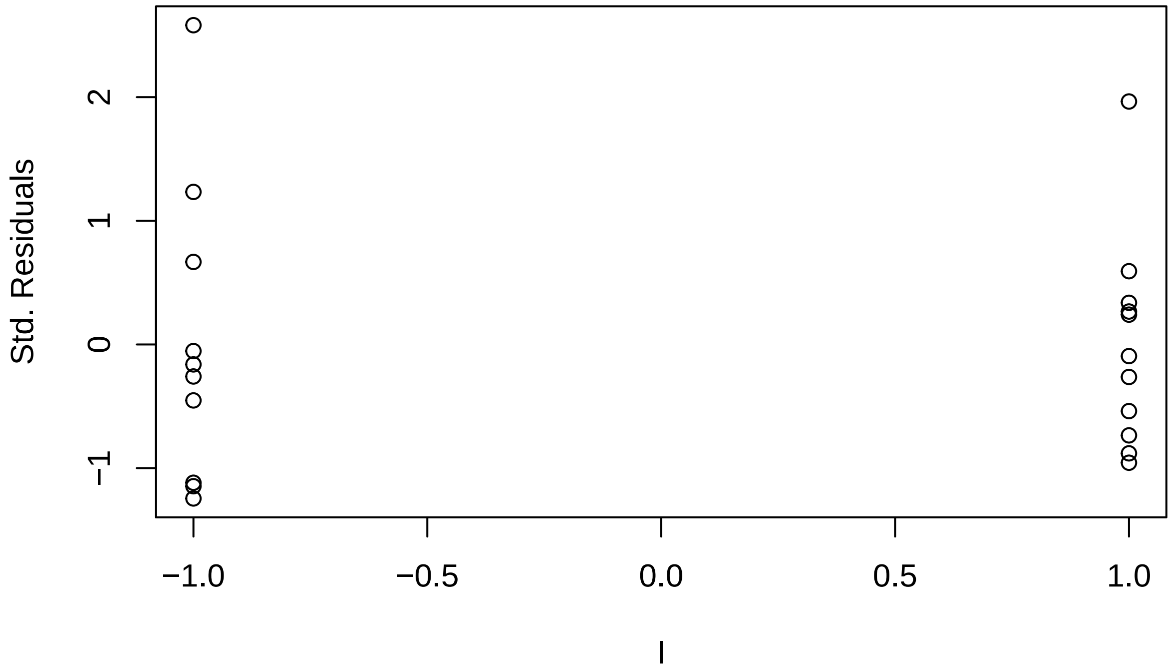
```r
qqnorm(elect_model_1.residuals, ylab="Sample Quantiles", xlab="Theoretical Quantiles")
qqline(elect_model_1.residuals)
```



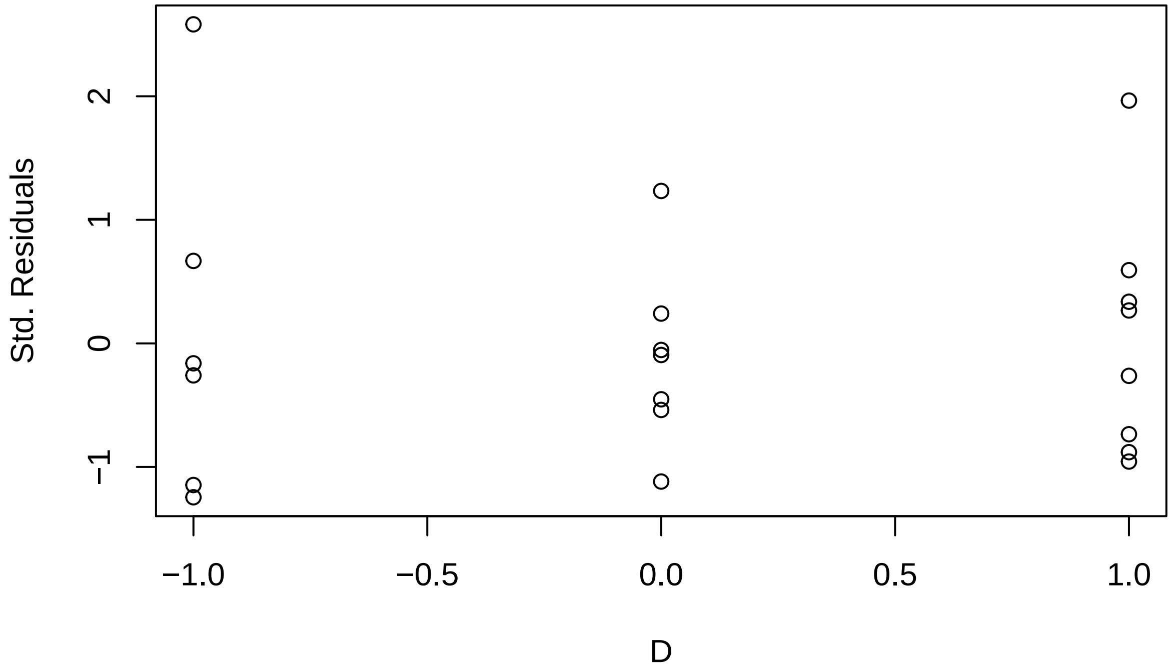
```
plot(fitted.values(elect_model_1), elect_model_1.residuals, ylab="Std. Residuals", xlab="Response")
```



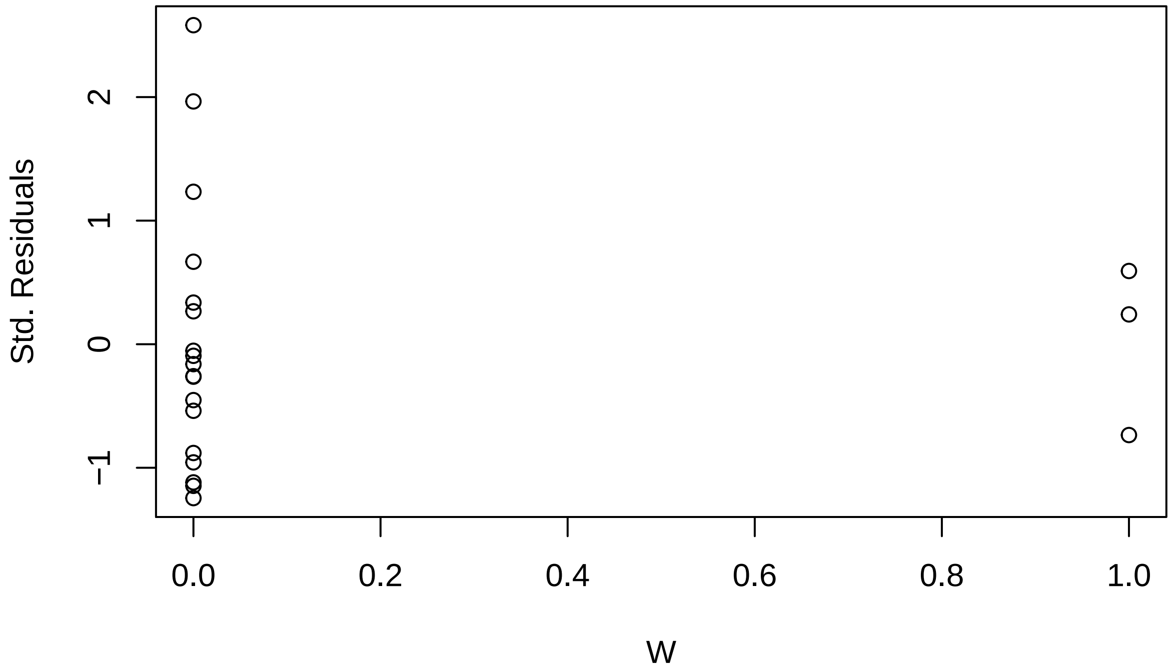
```
plot(elect_data_2a$I, elect_model_1.residuals, ylab="Std. Residuals", xlab="I")
```



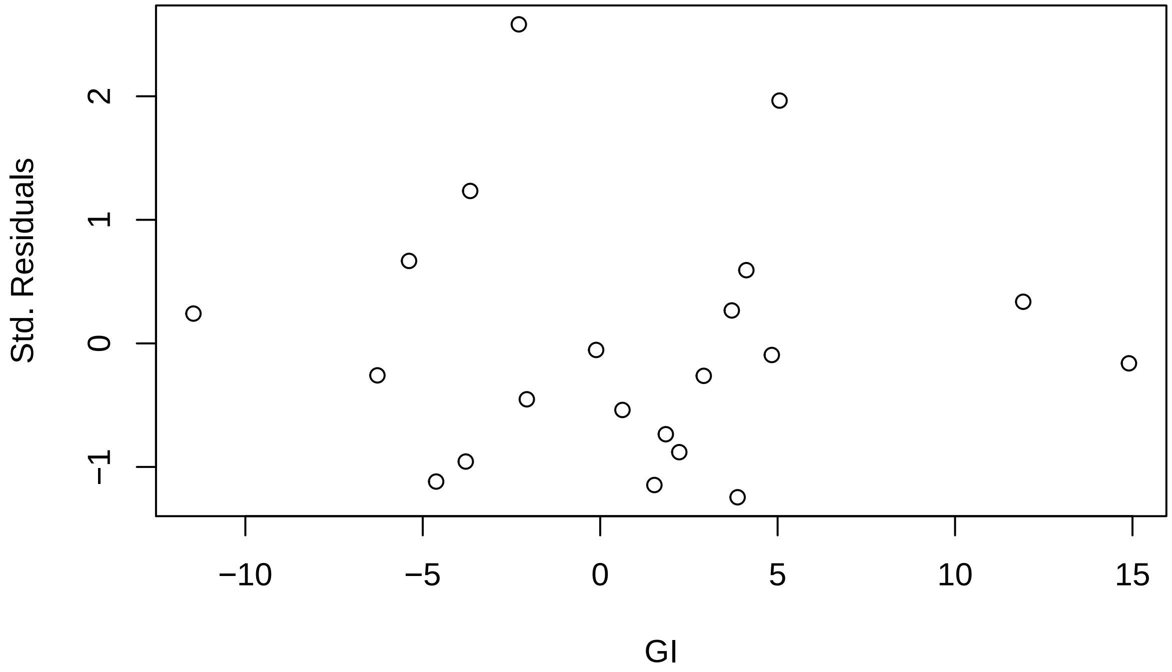
```r
plot(elect_data_2a$D, elect_model_1.residuals, ylab="Std. Residuals", xlab="D")
```



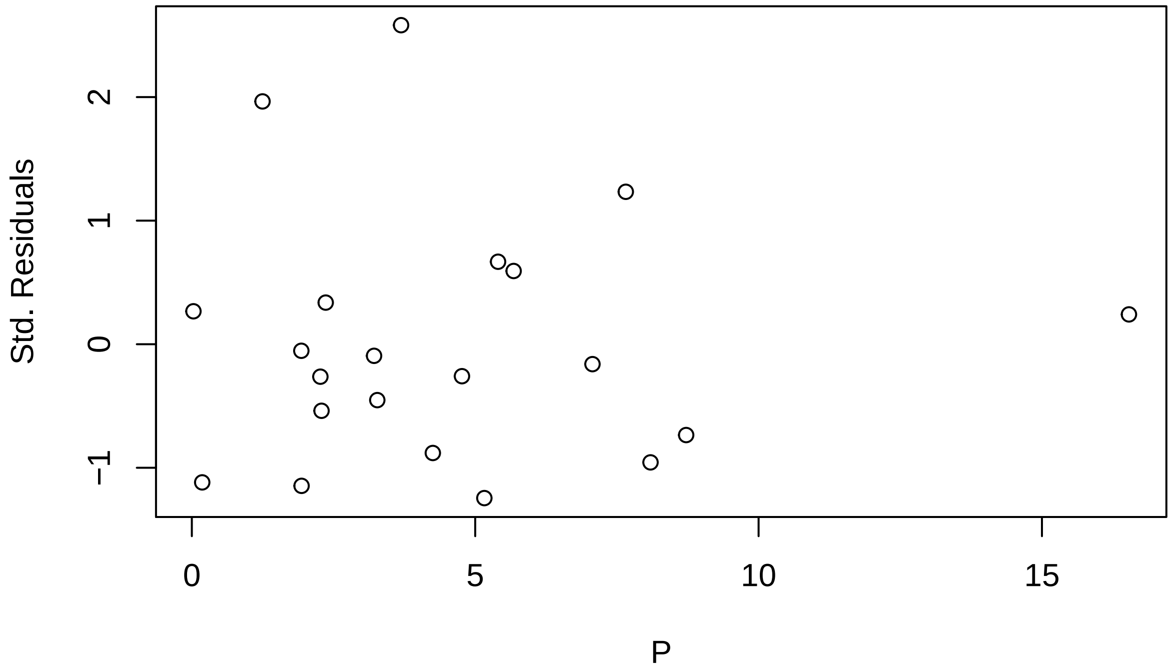
```r
plot(elect_data_2a$W, elect_model_1.residuals, ylab="Std. Residuals", xlab="W")
```



```
plot(elect_data_2a$GI, elect_model_1.residuals, ylab="Std. Residuals", xlab="GI")
```



```
plot(elect_data_2a$P, elect_model_1.residuals, ylab="Std. Residuals", xlab="P")
```

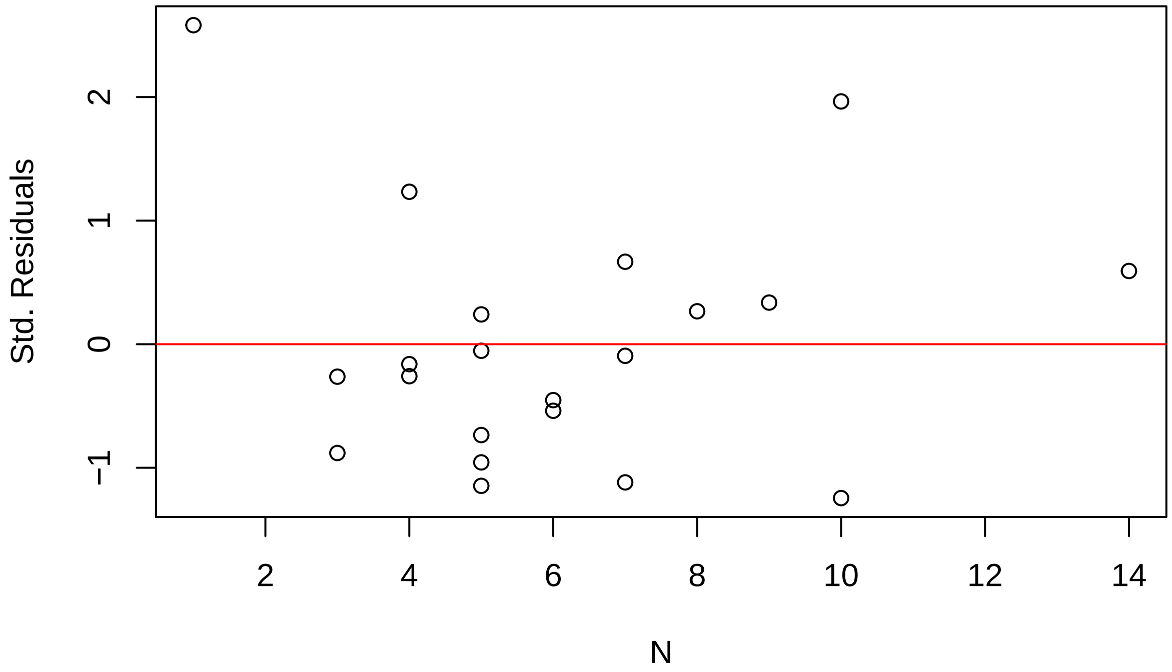


```
plot(elect_data_2a$N, elect_model_1.residuals, ylab="Std. Residuals", xlab="N")
abline(a=0, b=0, col="red")

cor(elect_data_2a)
```

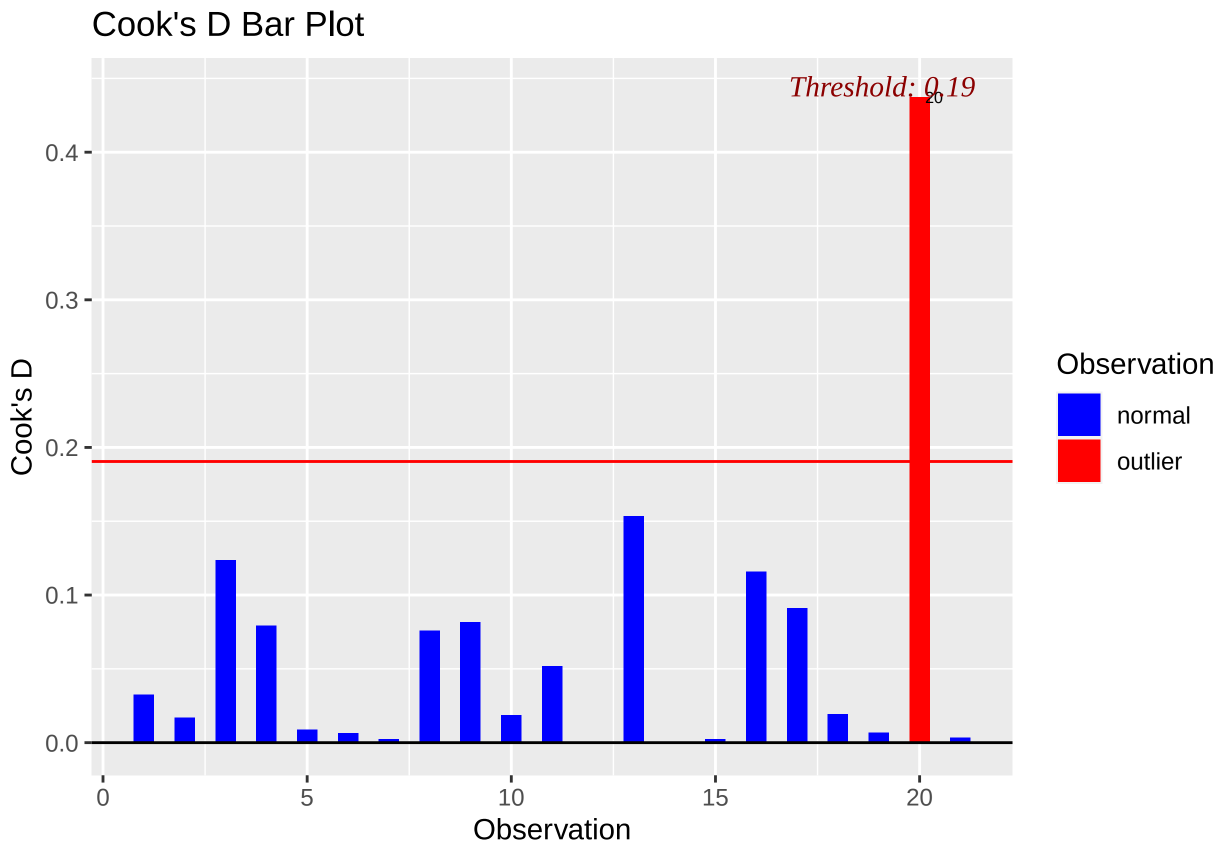
```
##              V         I           D           W          GI           P
## V   1.00000000 0.3464773  0.49916995 -0.09189944  0.7824082 -0.33232716
## I   0.34647728 1.0000000  0.81744307  0.38924947  0.2056066  0.11921266
## D   0.49916995 0.8174431  1.00000000  0.28767798  0.1999737 -0.07290826
## W  -0.09189944 0.3892495  0.28767798  1.00000000 -0.1868577  0.64831150
## GI  0.78240818 0.2056066  0.19997369 -0.18685769  1.0000000 -0.38056872
## P  -0.33232716 0.1192127 -0.07290826  0.64831150 -0.3805687  1.00000000
## N   0.14667760 0.2650860  0.28350827  0.27186362  0.2926510 -0.16705075
##             N
## V   0.1466776
## I   0.2650860
## D   0.2835083
## W   0.2718636
## GI  0.2926510
## P  -0.1670507
## N   1.0000000
```

```
library(olsrr)
```

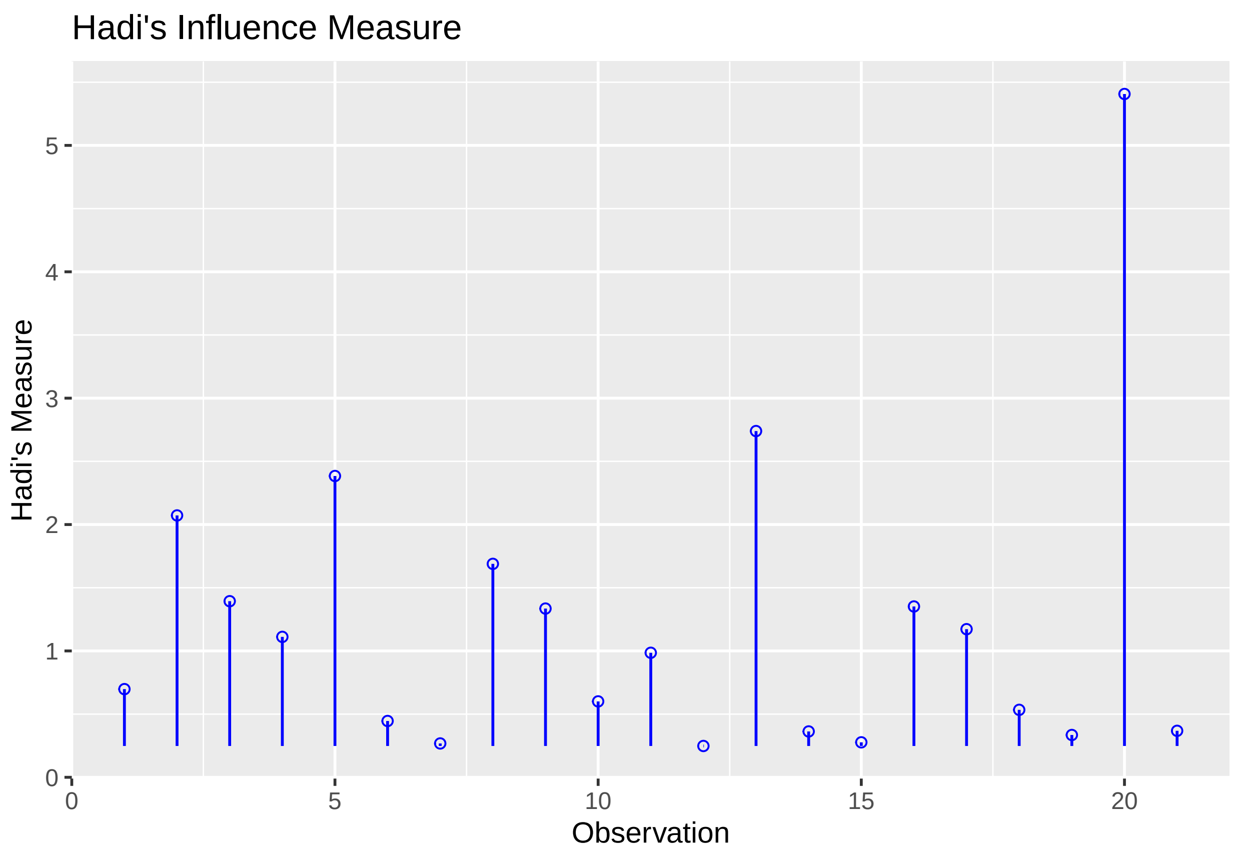
```
##
## Attaching package: 'olsrr'

## The following object is masked from 'package:datasets':
##
##     rivers
```



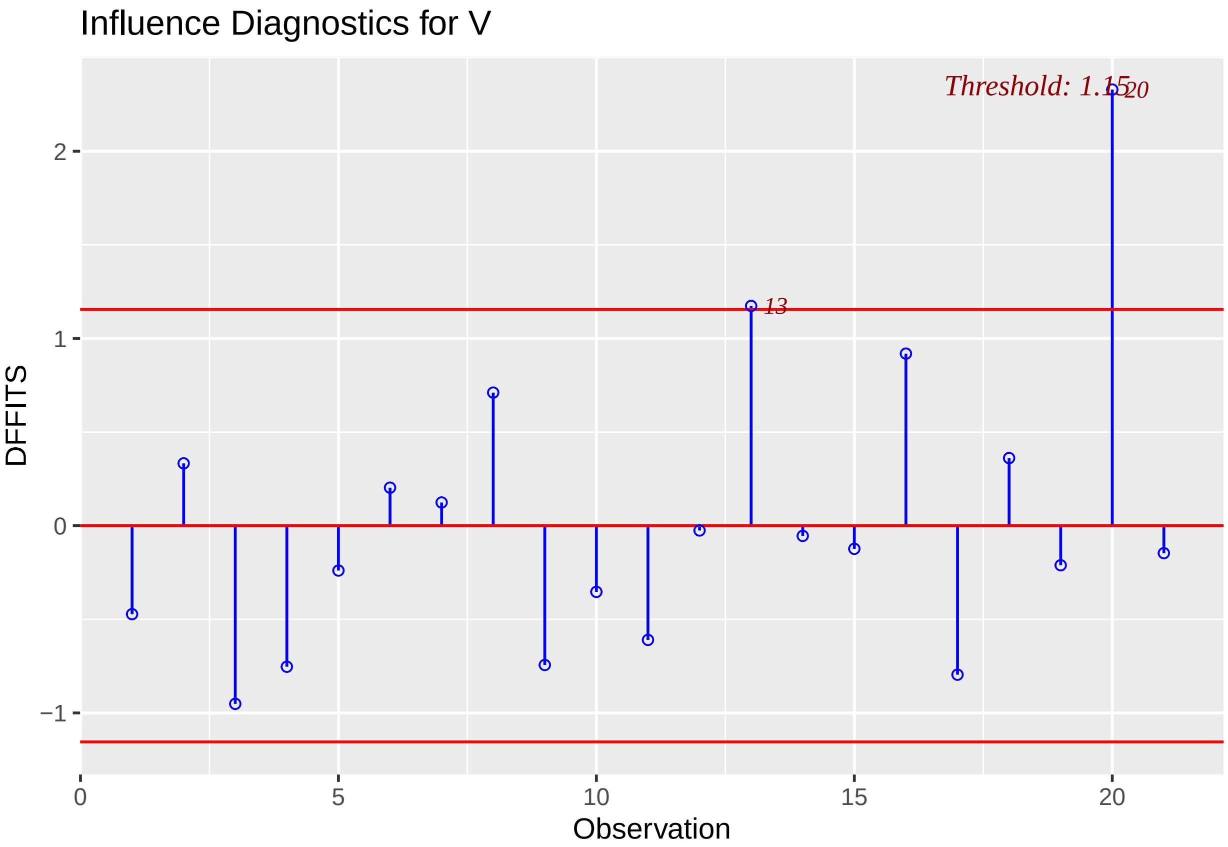
```
ols_plot_cooksd_bar(elect_model_1)
```



```
ols_plot_hadi(elect_model_1)
```



```
ols_plot_dffits(elect_model_1)
```



From the above residual-index and leverage-index plots, the leverages do not follow a random pattern, since the earlier observations appear to have more leverage. The standardised residuals largely follow a random pattern with the exception of observation 20 which has quite a high standardised residual. Hence (2.1) is likely violated.

Under the normality of errors assumption (2.2), the ordered residuals should be approximately form a straight line with slope 1 and intercept 0 with the quantiles of the standard normal distribution. From examining the Q-Q plot of the standardised residuals against the standard normal distribution, we notice that the sample quantiles mostly appear to match the theoretical quantiles with the exception of the last 3 points hence it is likely (2.2) is violated.

Assumption (2.3) is assumed to be true by least squares in an intercept model.

The standard residuals should be uncorrelated with the predictor variables/response. As such when plotting the std. residuals against each of the predictors, we expect to see a random scatter of points, which is true for I, D, Response, GI, P and N but NOT for W (although this can likely be forgiven since there are very few points where $W = 1$.) At every fitted value, the spread of the residuals is roughly the same with an outlying point or two in some of the plots. Hence the constant variance assumption (2.4) is satisfied.

Analysis of the std residuals vs predictors/fitted values plot shows no discernible pattern so it is likely that assumption (1) on linearity is satisfied.

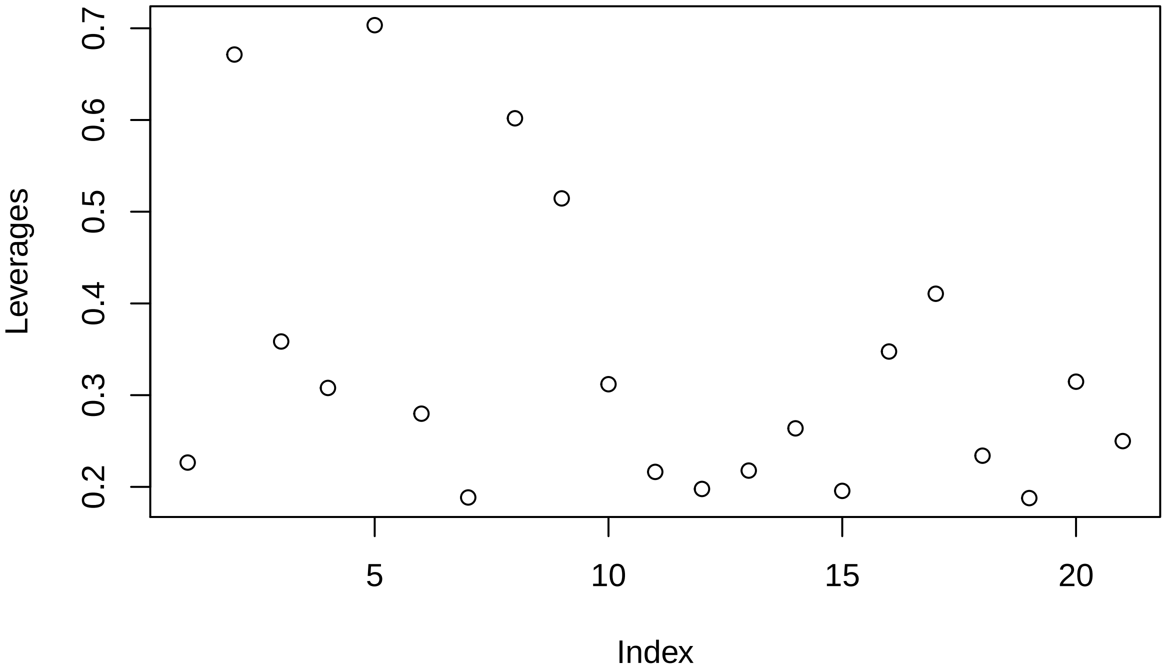
Of these assumptions, (3.2) and (3.1) are difficult to assume. Let us examine the independence of the predictors, i.e. assumption (3.3). Besides $Cor(D, I) = 0.817$ and $Cor(W, P) = 0.648$, the pairwise correlations between predictors is low. So (3.3) is likely satisfied.

By analyzing the Cook's distance, DFITS and Hadi's plots, we see that observations 13 and 20 is more influential than the rest, hence assumption (4) is violated.

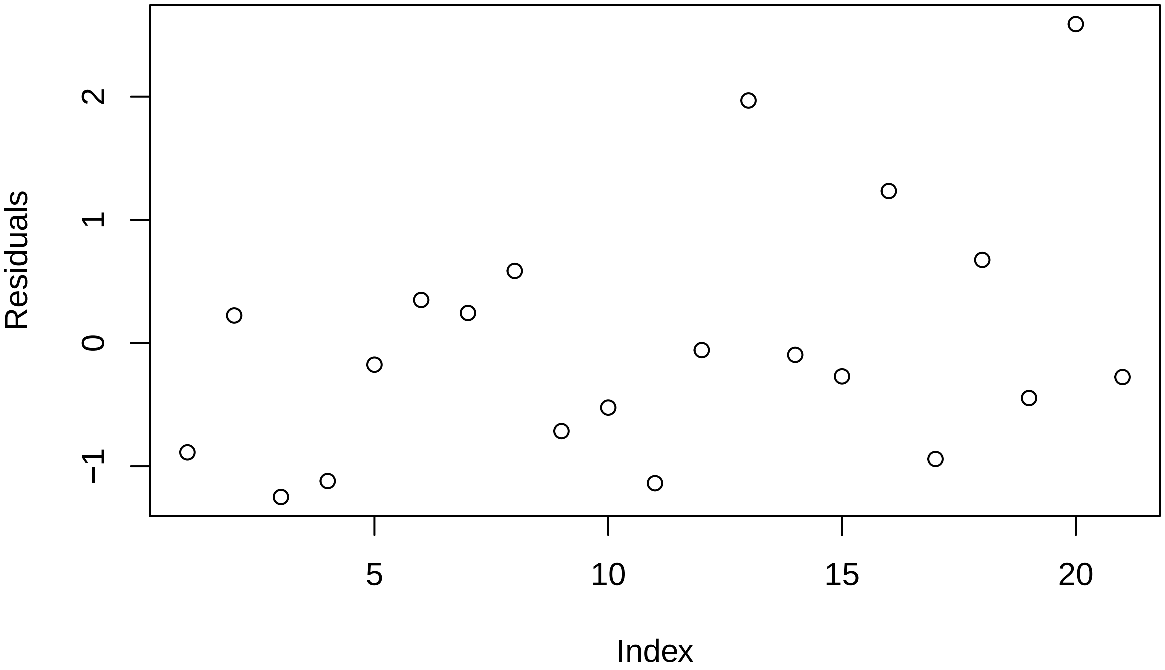
```
elect_data_5 <- subset(elect_data,select=c("V","I","D","W","GI","P","N"))
elect_data_5$Y <- log((elect_data$V)/(1-elect_data$V)) # by default is natural log
elect_data_5 <- subset(elect_data_5,select=c("Y","I","D","W","GI","P","N")) # drop V
elect_model_q5 = lm(Y ~ ., data = elect_data_5)
summary(elect_model_q5)
```

```
##
## Call:
## lm(formula = Y ~ ., data = elect_data_5)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.16746 -0.08279 -0.01588  0.04936  0.35640
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.043781   0.130199   0.336   0.7417
## I           -0.082056   0.068328  -1.201   0.2497
## D            0.222161   0.083178   2.671   0.0183 *
## W            0.050279   0.170896   0.294   0.7729
## GI           0.039359   0.007162   5.496 7.88e-05 ***
## P           -0.002952   0.016193  -0.182   0.8580
## N           -0.020706   0.015399  -1.345   0.2001
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1663 on 14 degrees of freedom
## Multiple R-squared:  0.7917, Adjusted R-squared:  0.7025
## F-statistic:  8.87 on 6 and 14 DF,  p-value: 0.0004095
```

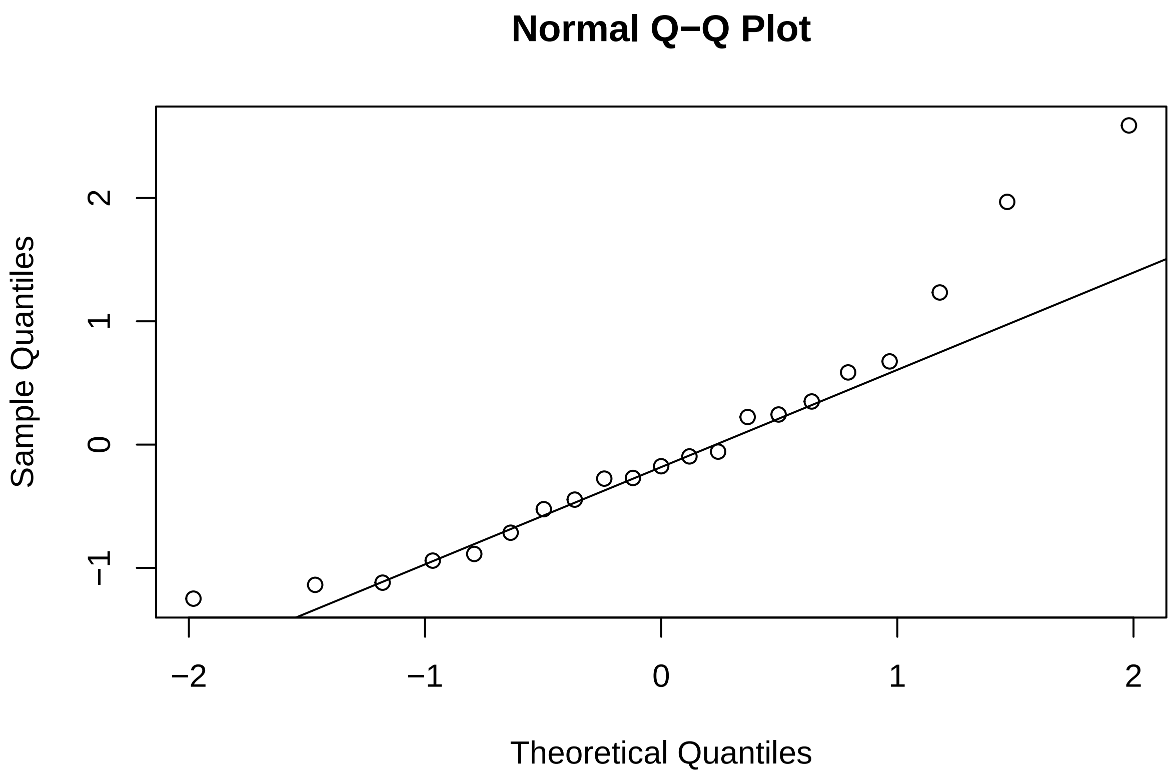
```
leverages <- hatvalues(elect_model_q5) # compute leverages for model Y
elect_model_q5.residuals <- rstandard(elect_model_q5)

plot(seq(1,length(elect_data$V)), leverages, xlab="Index", ylab="Leverages")
```



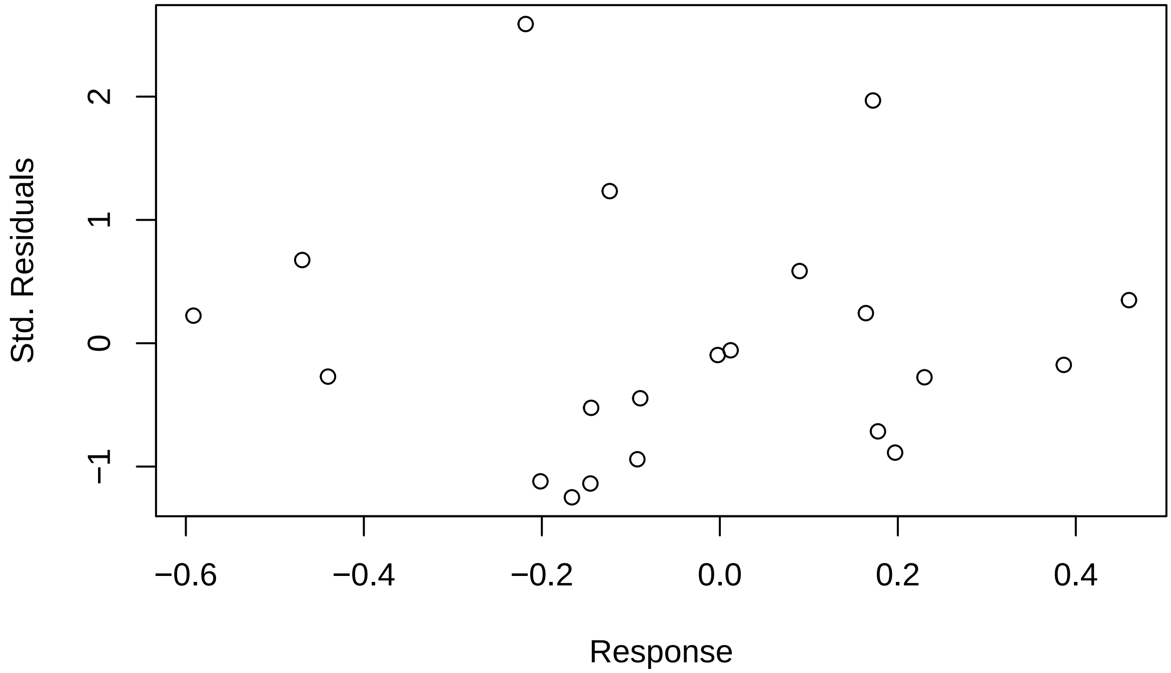
```
plot(seq(1,length(elect_data$V)), elect_model_q5.residuals, xlab="Index", ylab="Residuals")
```



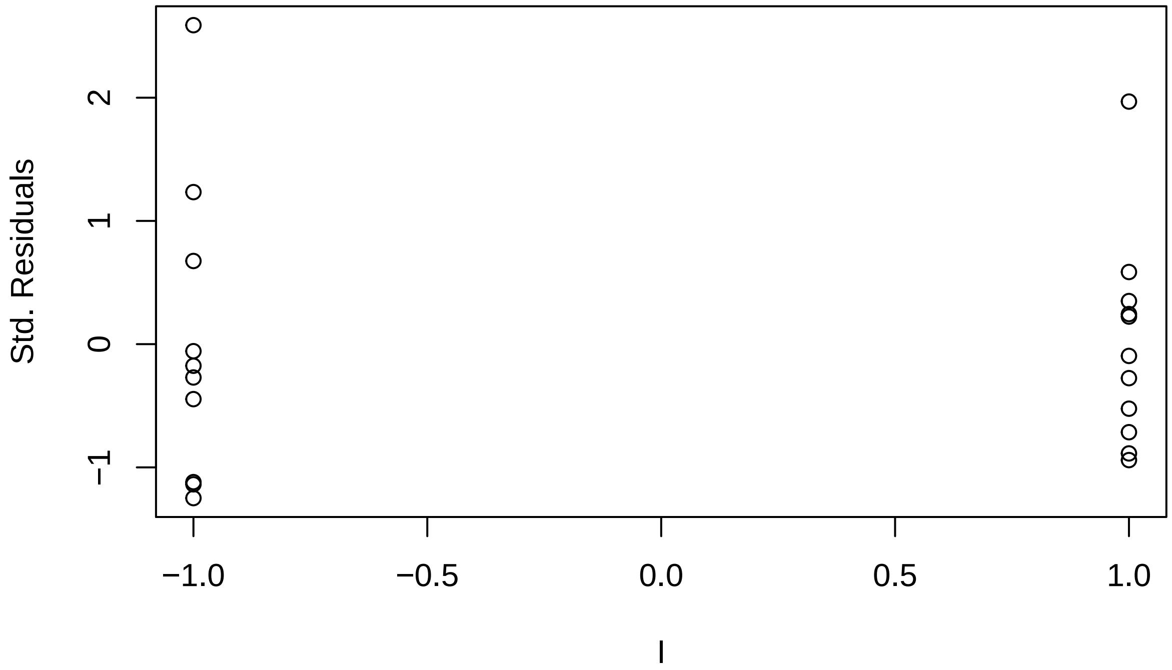
```
qqnorm(elect_model_q5.residuals, ylab="Sample Quantiles", xlab="Theoretical Quantiles")
qqline(elect_model_q5.residuals)
```



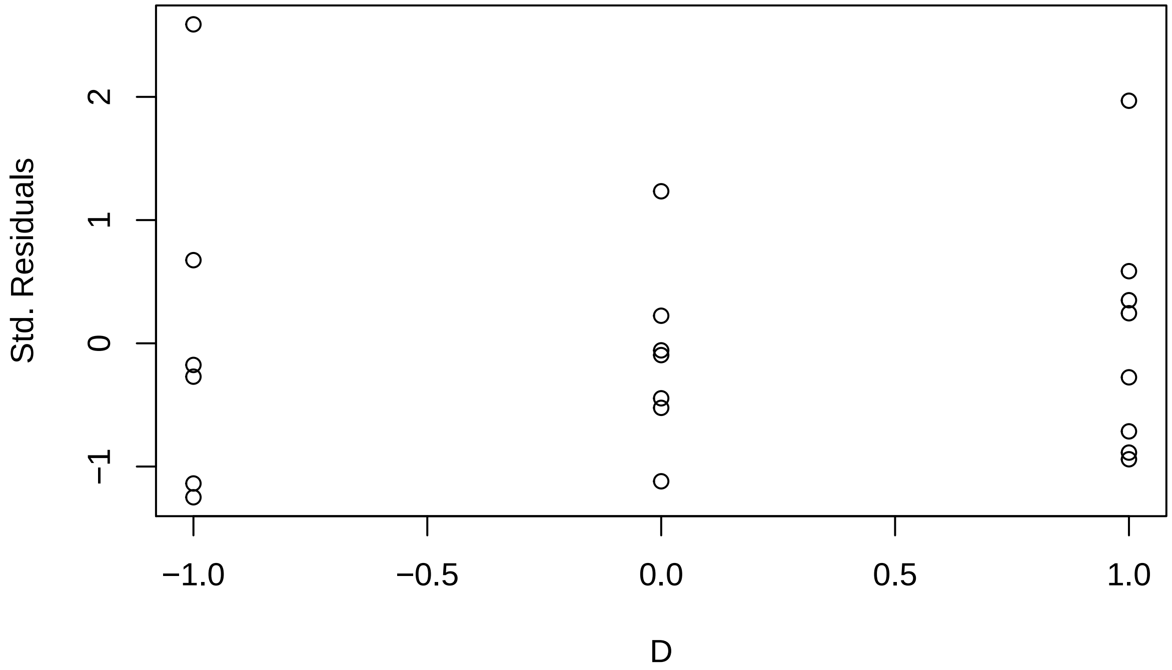
```
plot(fitted.values(elect_model_q5), elect_model_q5.residuals, ylab="Std. Residuals", xlab="Response")
```



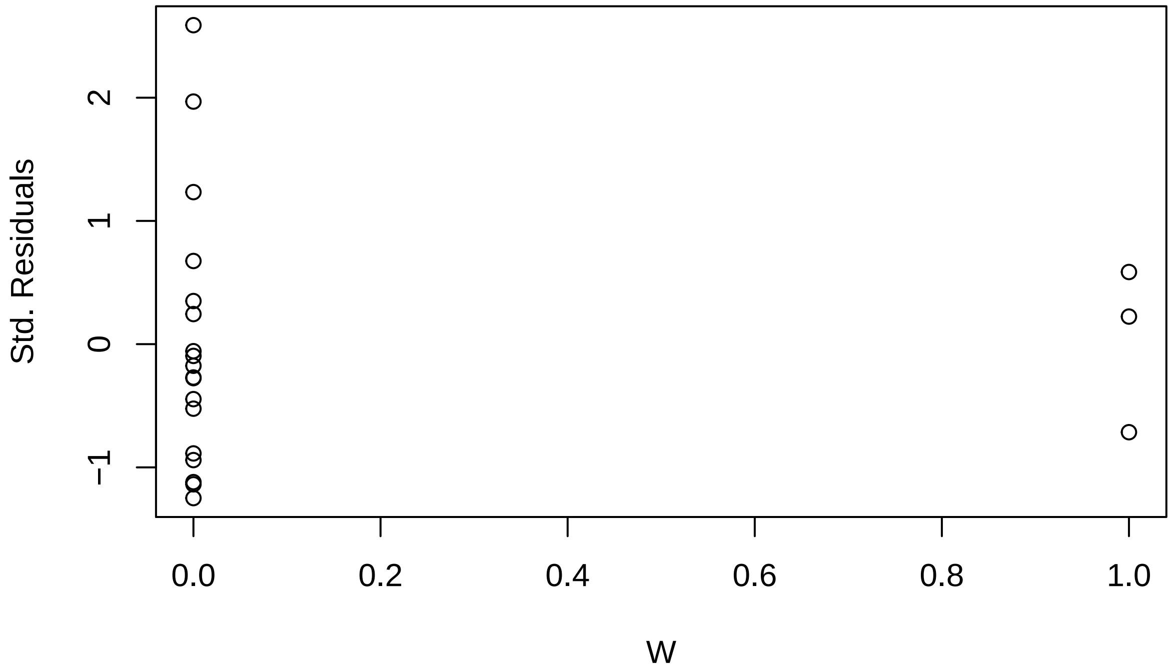
```
plot(elect_data_2a$I, elect_model_q5.residuals, ylab="Std. Residuals", xlab="I")
```



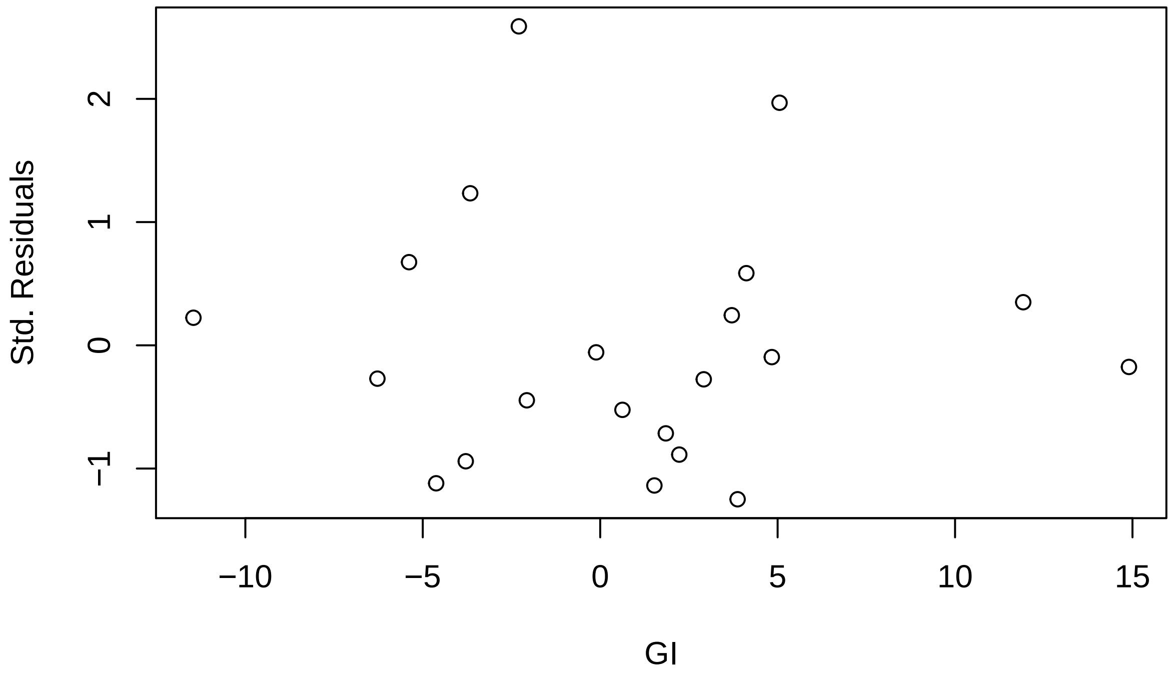
```
plot(elect_data_2a$D, elect_model_q5.residuals, ylab="Std. Residuals", xlab="D")
```



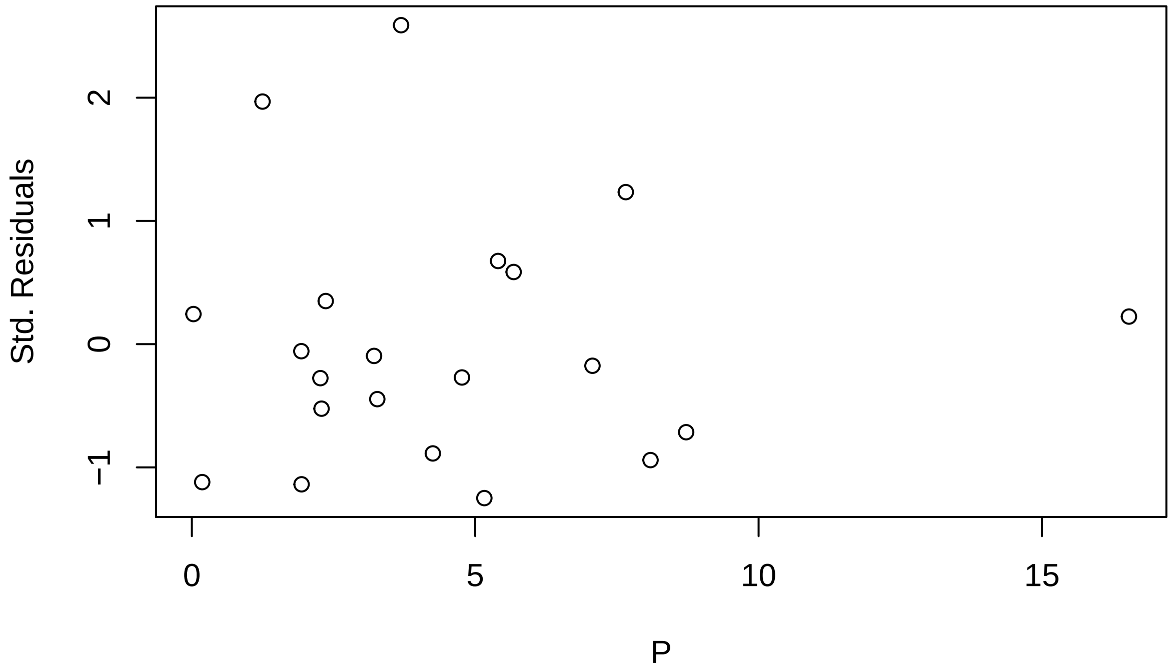
```
plot(elect_data_2a$W, elect_model_q5.residuals, ylab="Std. Residuals", xlab="W")
```



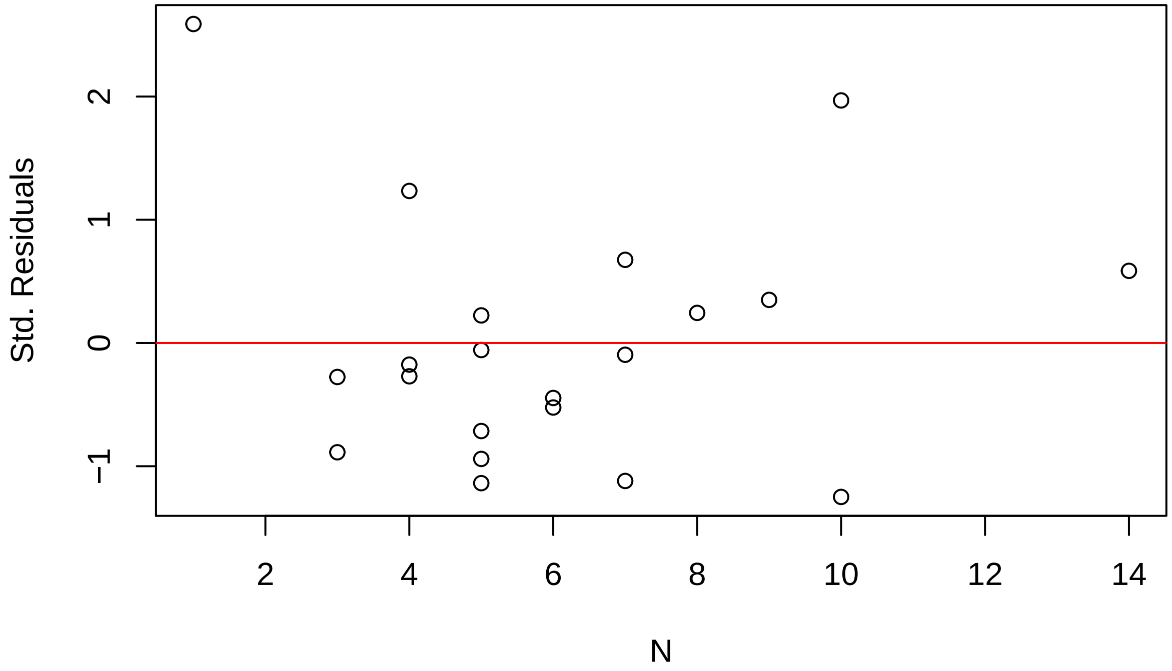
```
plot(elect_data_2a$GI, elect_model_q5.residuals, ylab="Std. Residuals", xlab="GI")
```



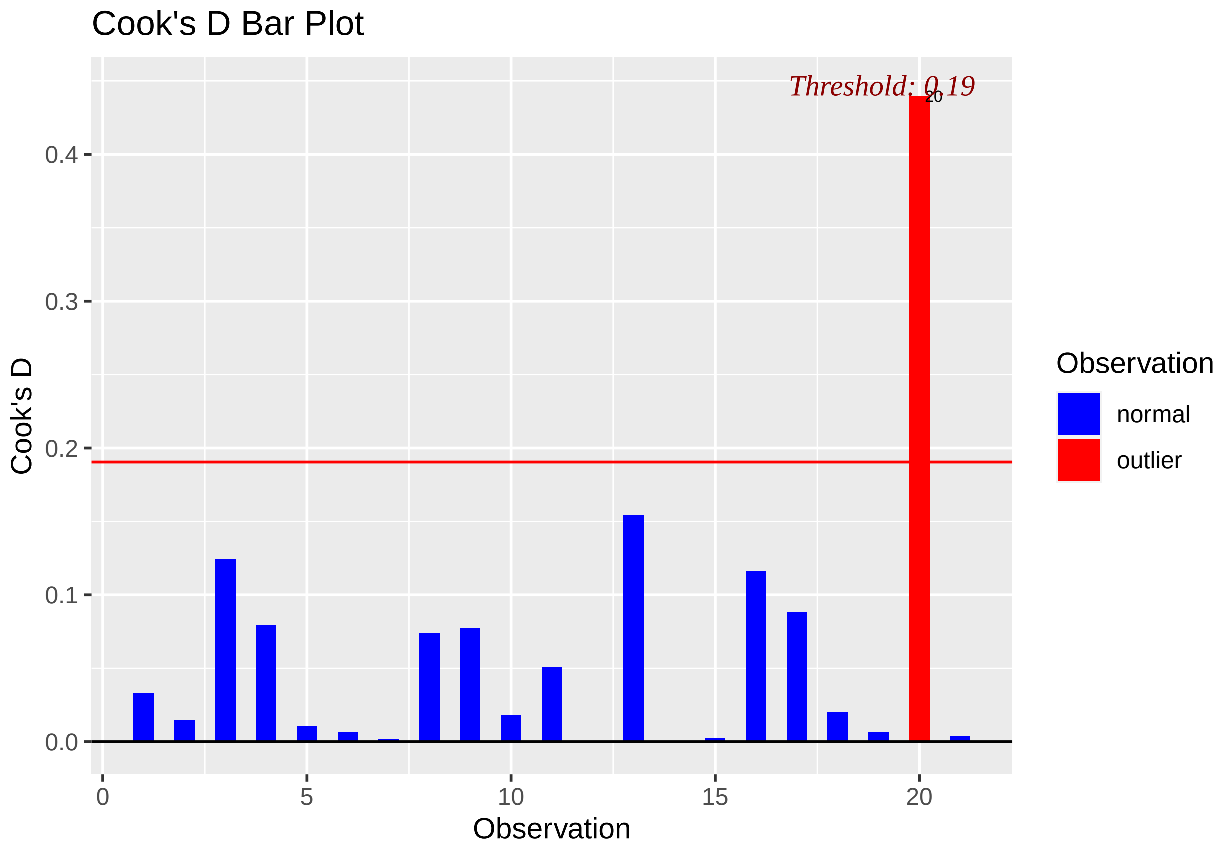
```
plot(elect_data_2a$P, elect_model_q5.residuals, ylab="Std. Residuals", xlab="P")
```



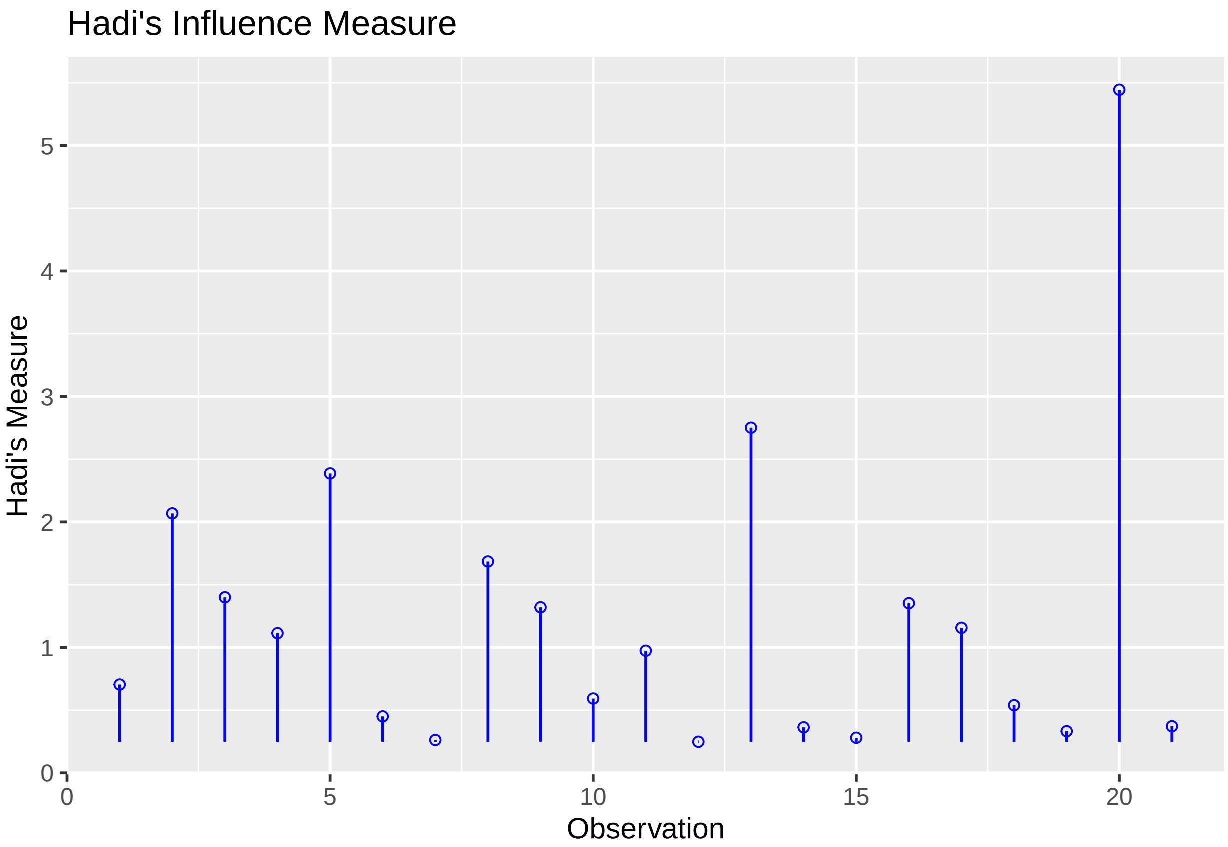
```
plot(elect_data_2a$N, elect_model_q5.residuals, ylab="Std. Residuals", xlab="N")
abline(a=0, b=0, col="red")
```



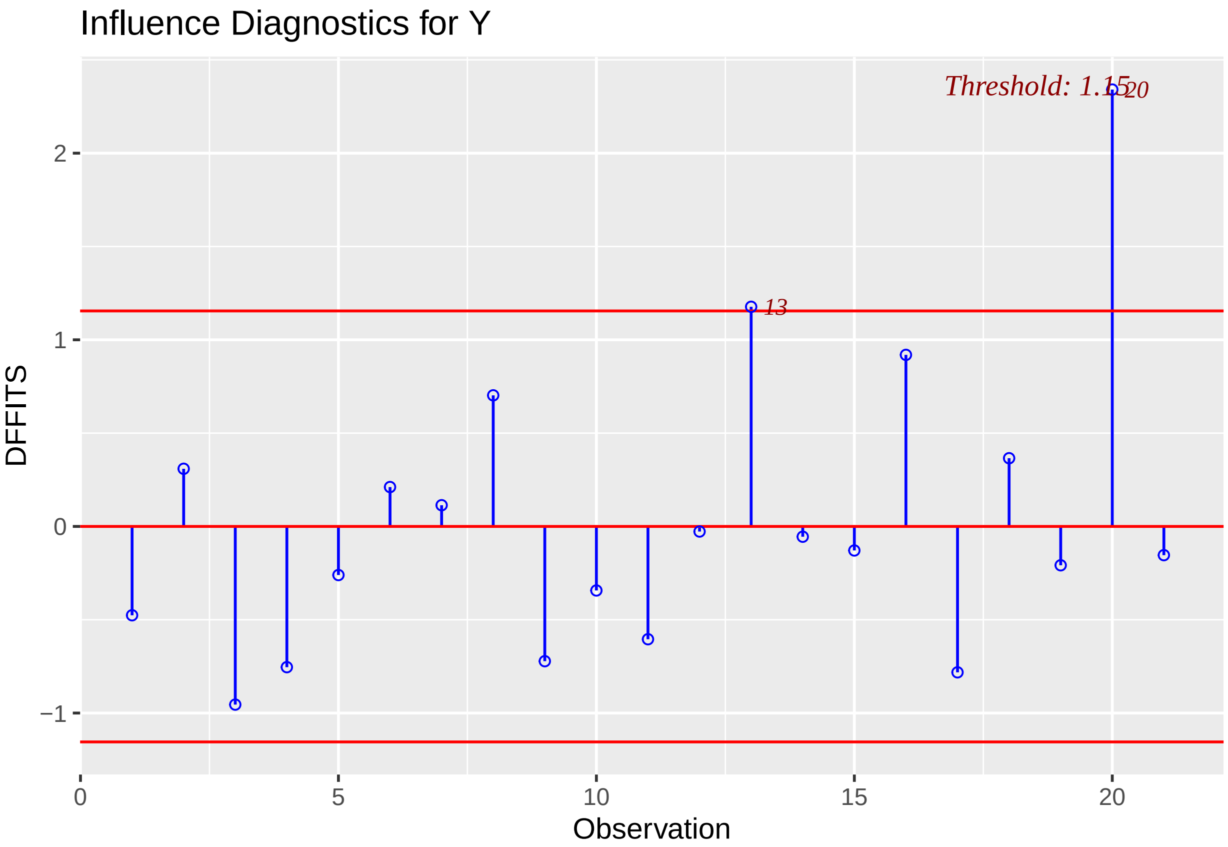
```
ols_plot_cooksd_bar(elect_model_q5)
```



```
ols_plot_hadi(elect_model_q5)
```



```
ols_plot_dffits(elect_model_q5)
```



From the above residual-index and leverage-index plots, the leverages do not follow a random pattern, since the earlier observations appear to have more leverage. The standardised residuals largely follow a random pattern with the exception of observation 20 which has quite a high standardised residual. Hence (2.1) is likely violated.

Under the normality of errors assumption (2.2), the ordered residuals should be approximately form a straight line with slope 1 and intercept 0 with the quantiles of the standard normal distribution. From examining the Q-Q plot of the standardised residuals against the standard normal distribution, we notice that the sample quantiles mostly appear to match the theoretical quantiles with the exception of the last 3 points, so again it is likely (2.2) is not satisfied.

Assumption (2.3) is assumed to be true by least squares in an intercept model.

The standard residuals should be uncorrelated with the predictor variables/response. As such when plotting the std. residuals against each of the predictors, we expect to see a random scatter of points, which is true for I, D, Response, GI, P and N but NOT for W (although this can likely be forgiven since there are very few points where $W = 1$.) At every fitted value, the spread of the residuals is roughly the same with an outlying point or two in some of the plots. Hence the constant variance assumption (2.4) is satisfied.

Analysis of the std residuals vs predictors/fitted values plot shows no discernible pattern so it is likely that assumption (1) on linearity is satisfied.

Of these assumptions, (3.2) and (3.1) are difficult to assume. Let us examine the independence of the predictors, i.e. assumption (3.3). Same as with the original model, besides $Cor(D, I) = 0.817$ and $Cor(W, P) = 0.648$, the pairwise correlations between predictors is low. So (3.3) is likely satisfied.

By analyzing the Cook's distance, DFITS and Hadi's plots, we see that observations 13 and 20 are more influential than the rest, hence assumption (4) is violated.

In summary, the diagnostic plots are largely the same since Y and V are almost linearly related over the observed value range of V in the existing dataset.

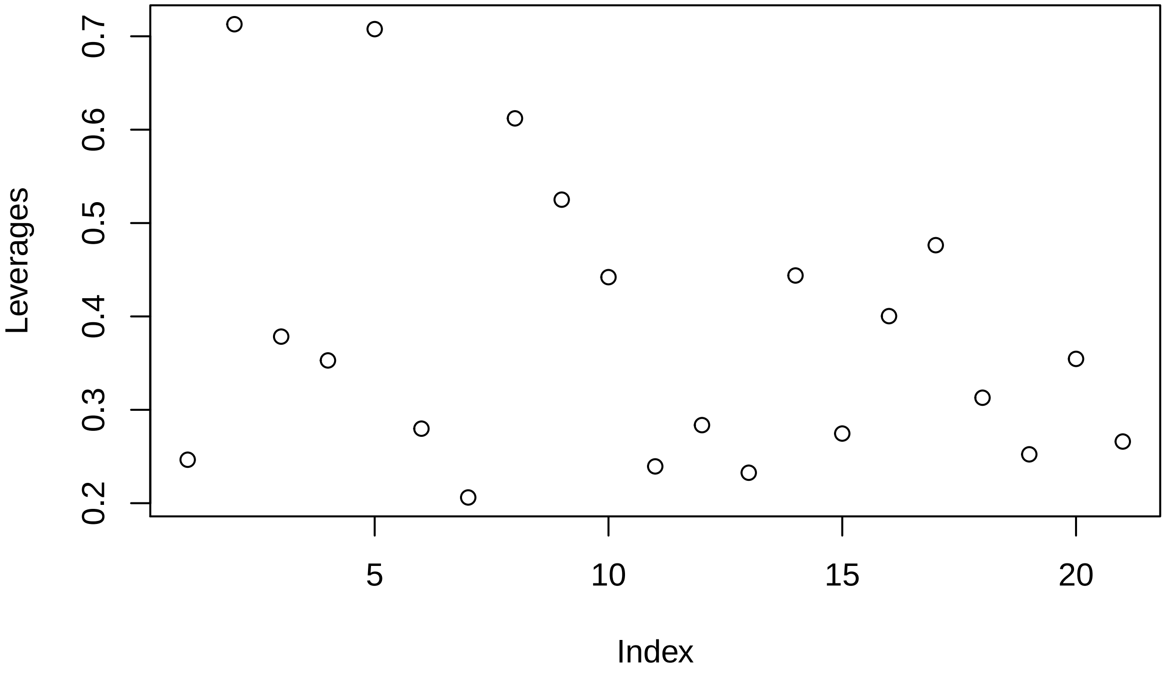
Q5b) $Y = log(V/(1-V)) \implies e^Y = V/(1-V) \implies V = e^Y/(1+e^Y)$

Since $Y = \beta_0 + \beta_1 I + \beta_2 D + \beta_3 W + \beta_4(G \cdot I) + \beta_5 P + \beta_6 N + \epsilon$, the function is $V = f(Y) = \dfrac{e^Y}{1+e^Y}$ i.e. the logistic function.)

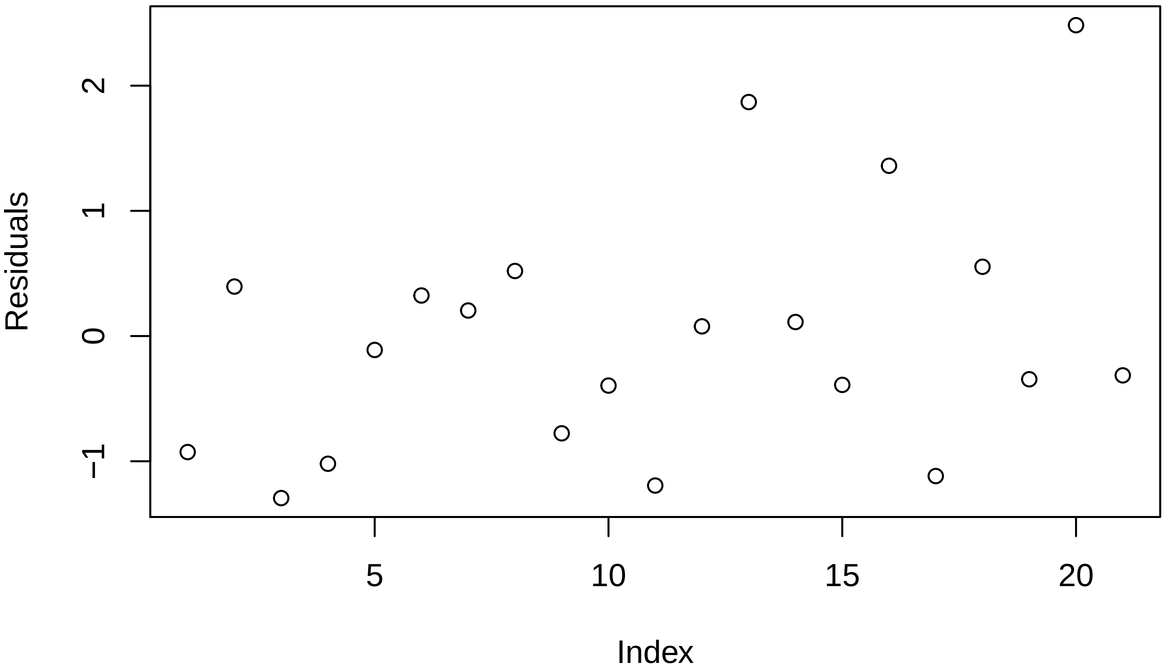
Q6

Q6a)

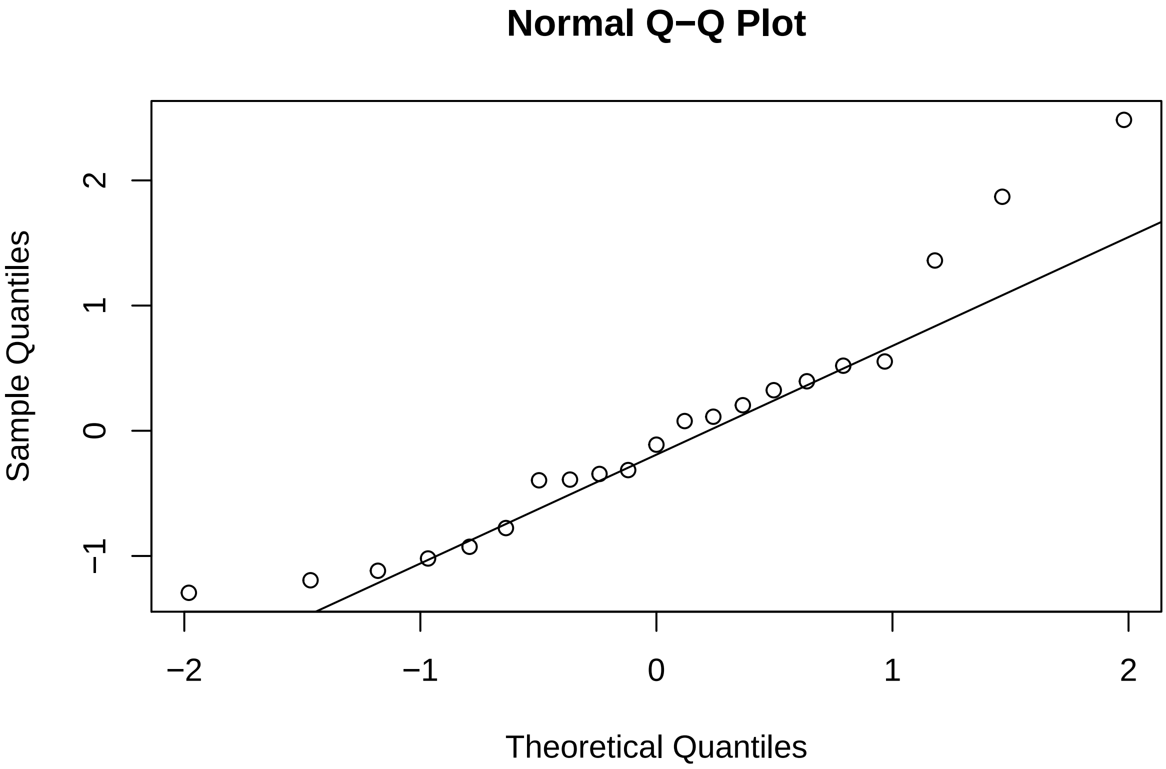
```
leverages <- hatvalues(elect_model_q3) # compute leverages for model V
elect_model_q3.residuals <- rstandard(elect_model_q3)

plot(seq(1,length(elect_data$V)), leverages, xlab="Index", ylab="Leverages")
```



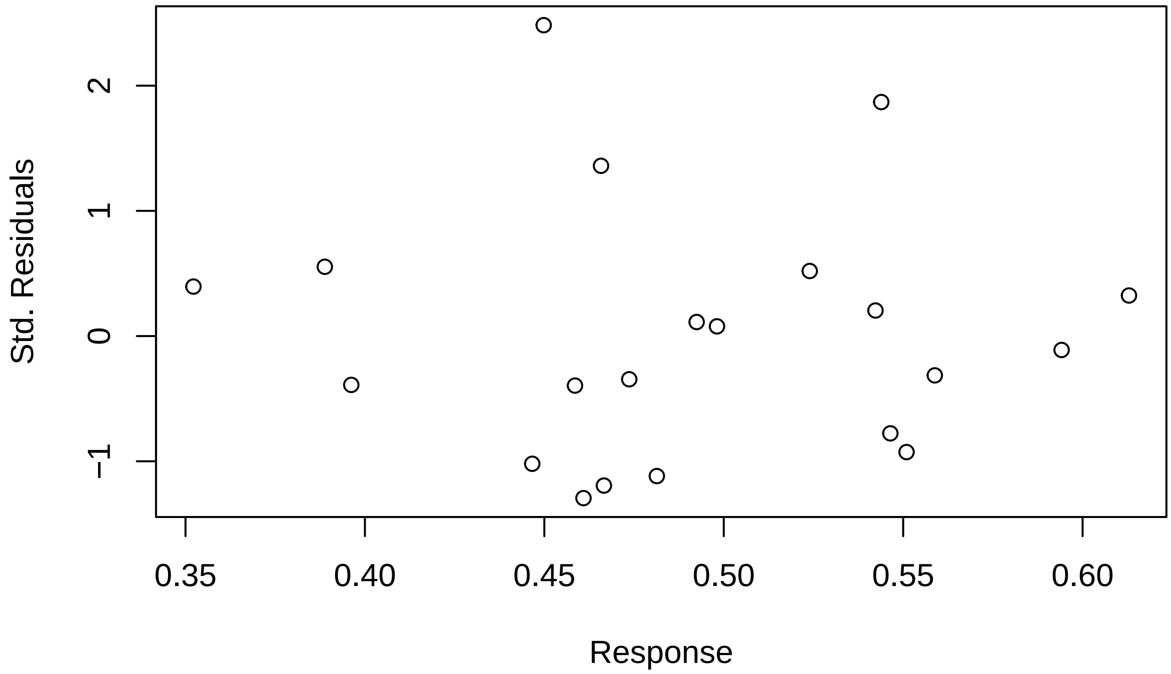
```
plot(seq(1,length(elect_data$V)), elect_model_q3.residuals, xlab="Index", ylab="Residuals")
```



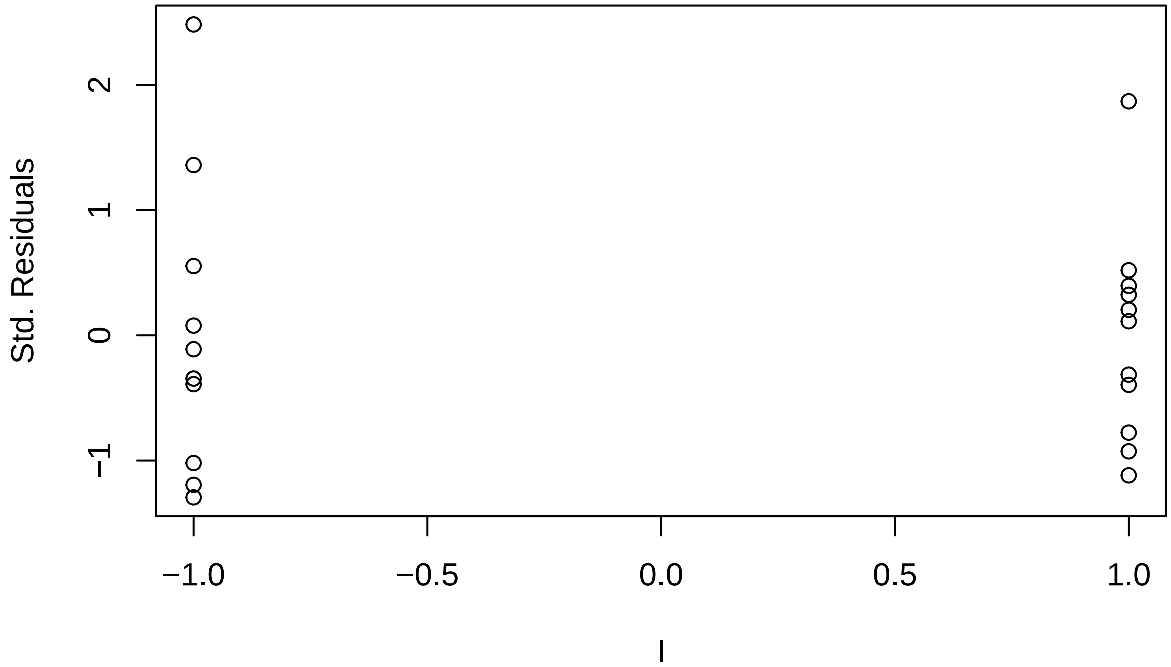
```
qqnorm(elect_model_q3.residuals, ylab="Sample Quantiles", xlab="Theoretical Quantiles")
qqline(elect_model_q3.residuals)
```



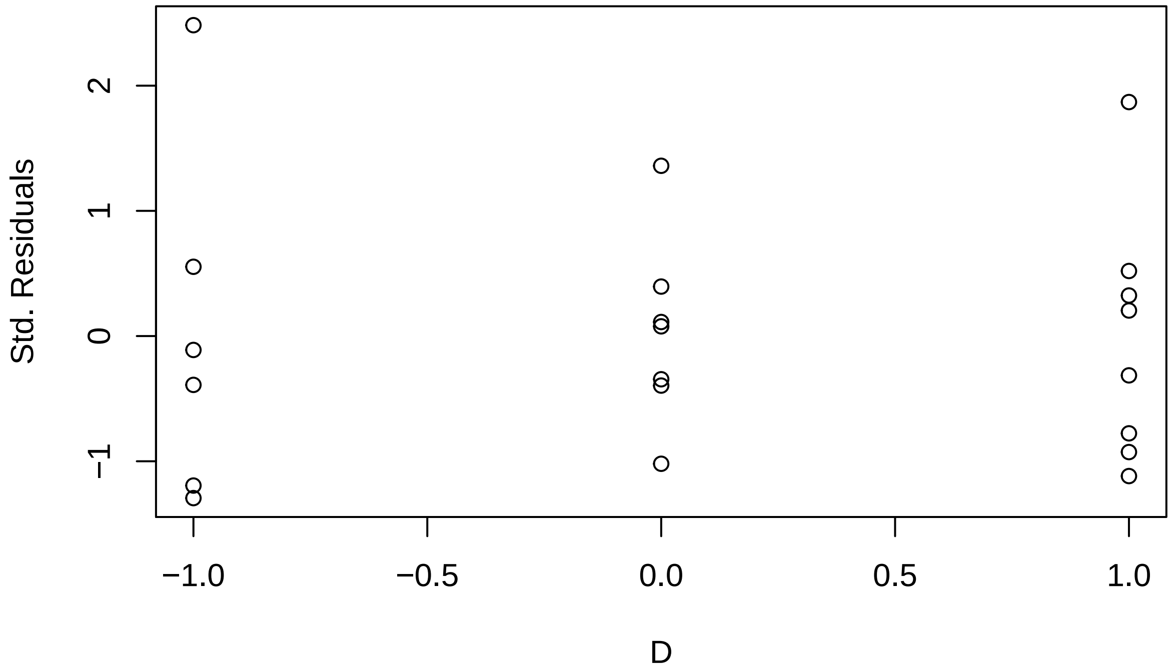
```
plot(fitted.values(elect_model_q3), elect_model_q3.residuals, ylab="Std. Residuals", xlab="Response")
```



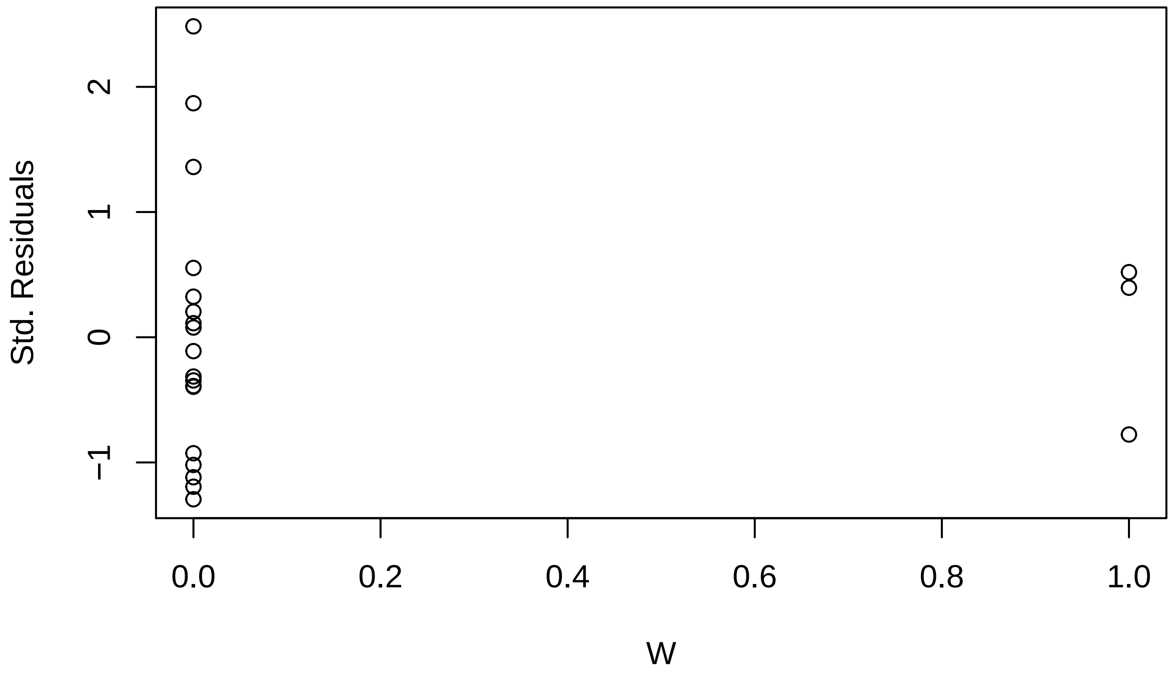
```
plot(elect_data_2a$I, elect_model_q3.residuals, ylab="Std. Residuals", xlab="I")
```



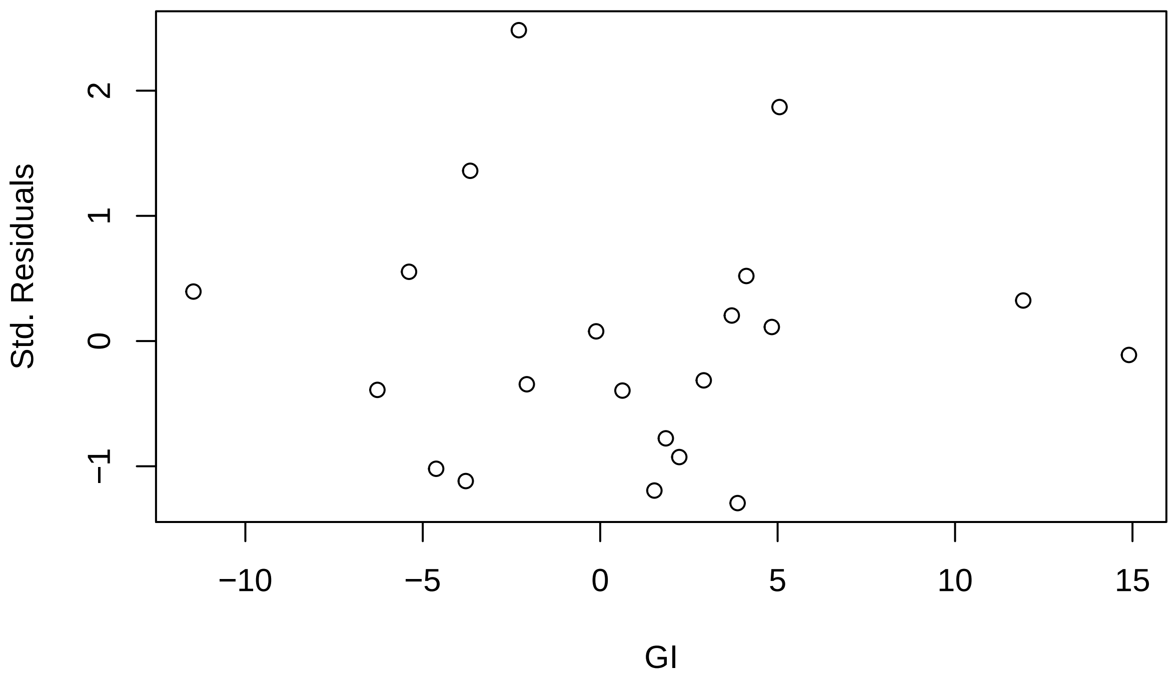
```
plot(elect_data_2a$D, elect_model_q3.residuals, ylab="Std. Residuals", xlab="D")
```



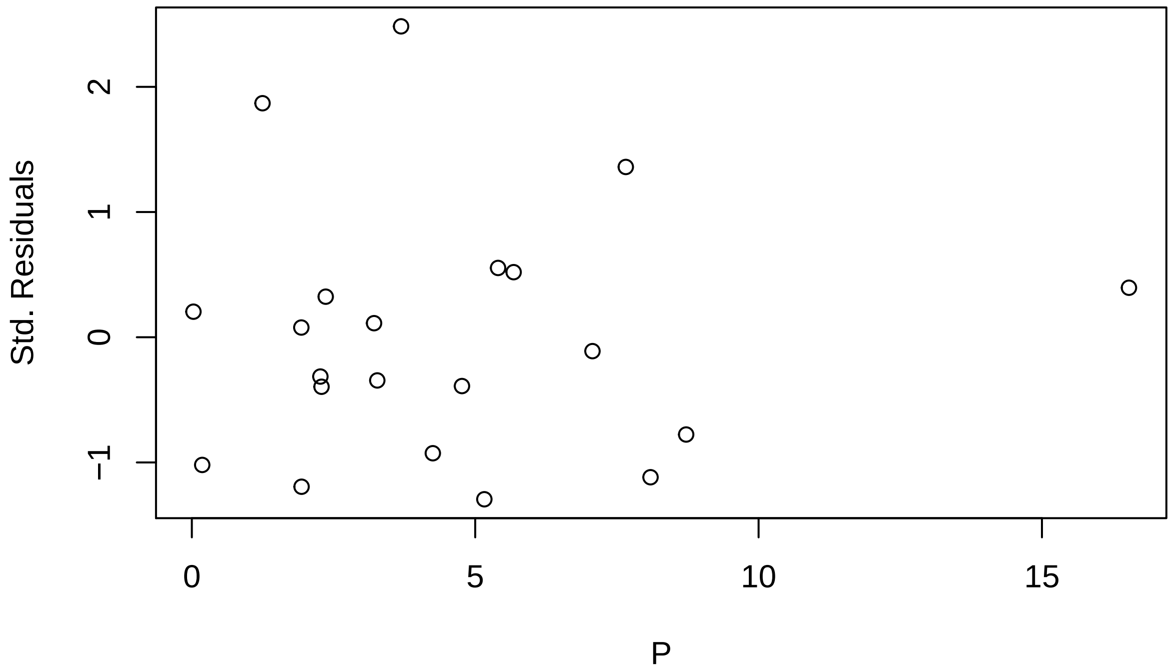
```
plot(elect_data_2a$W, elect_model_q3.residuals, ylab="Std. Residuals", xlab="W")
```



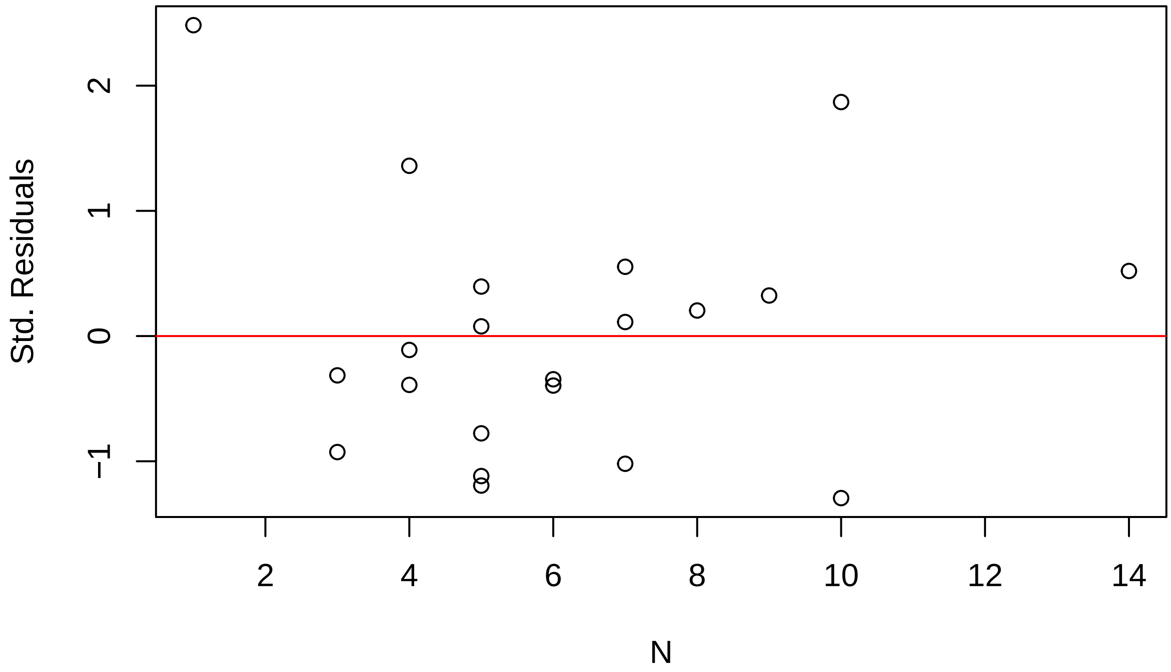
```
plot(elect_data_2a$GI, elect_model_q3.residuals, ylab="Std. Residuals", xlab="GI")
```



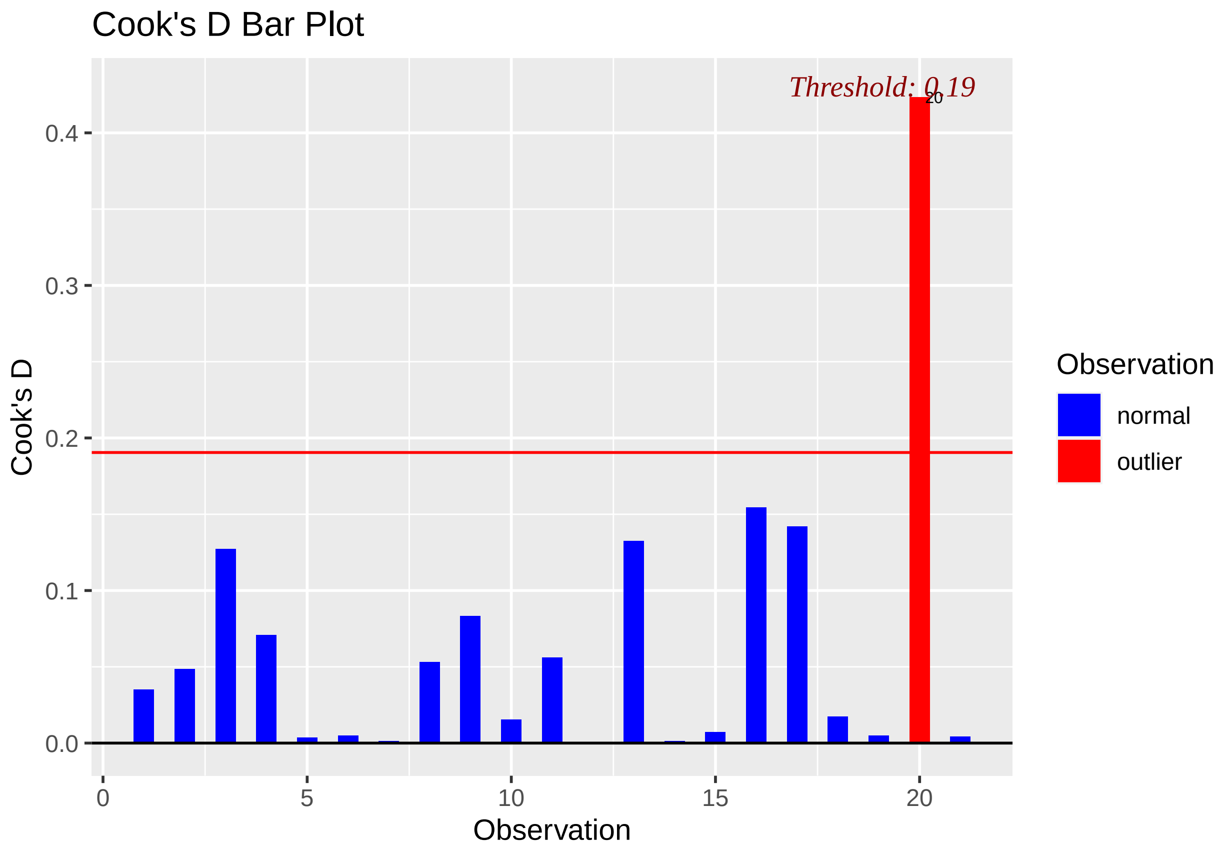
```
plot(elect_data_2a$P, elect_model_q3.residuals, ylab="Std. Residuals", xlab="P")
```



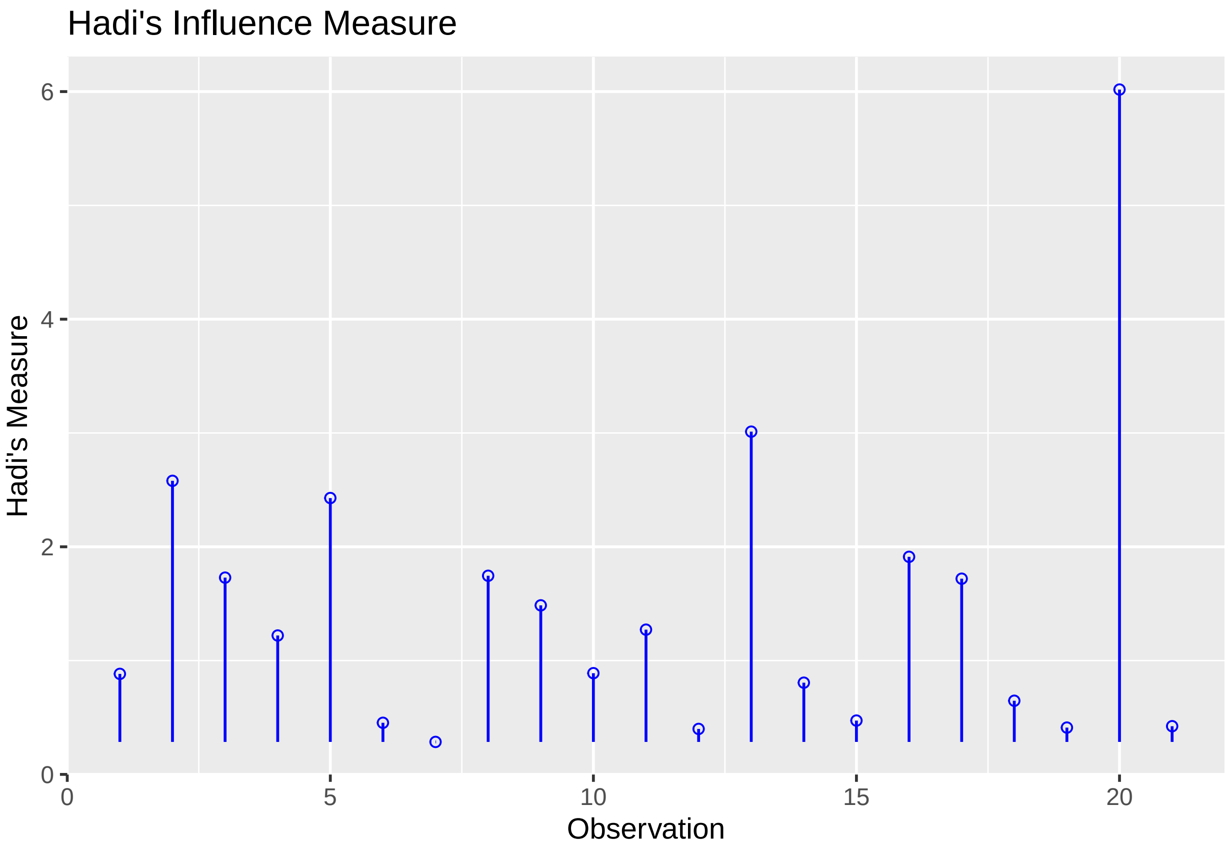
```
plot(elect_data_2a$N, elect_model_q3.residuals, ylab="Std. Residuals", xlab="N")
abline(a=0, b=0, col="red")
```



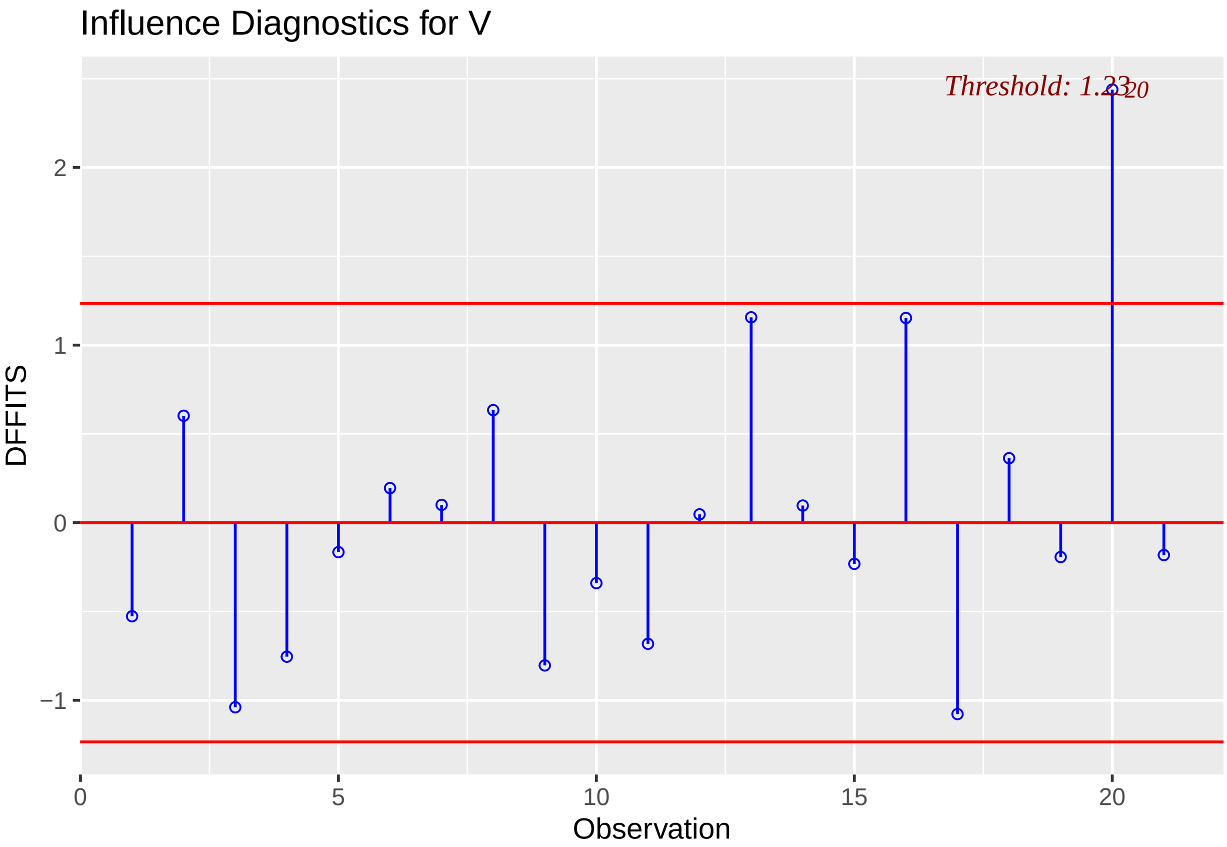
```
ols_plot_cooksd_bar(elect_model_q3)
```



```
ols_plot_hadi(elect_model_q3)
```



```
ols_plot_dffits(elect_model_q3)
```



```
elect_data_3c$D2 <- elect_data_3c$D2minusD1 + elect_data_3c$D1
elect_data_q6_ind <- subset(elect_data_3c,select=c("V","I","D1","D2","W","GI","P","N"))
cor(elect_data_q6_ind)
```

```
##              V          I         D1          D2           W          GI
## V   1.00000000  0.3464773  0.5823766 -0.27000143 -0.09189944  0.78240818
## I   0.34647728  1.0000000  0.7479576 -0.66332496  0.38924947  0.20560659
## D1  0.58237662  0.7479576  1.0000000 -0.49613894  0.24019223  0.35454668
## D2 -0.27000143 -0.6633250 -0.4961389  1.00000000 -0.25819889  0.02216247
## W  -0.09189944  0.3892495  0.2401922 -0.25819889  1.00000000 -0.18685769
## GI  0.78240818  0.2056066  0.3545467  0.02216247 -0.18685769  1.00000000
## P  -0.33232716  0.1192127 -0.1036814  0.01942015  0.64831150 -0.38056872
## N   0.14667760  0.2650860  0.2824205 -0.20532004  0.27186362  0.29265096
##              P          N
## V  -0.33232716  0.1466776
## I   0.11921266  0.2650860
## D1 -0.10368142  0.2824205
## D2  0.01942015 -0.2053200
## W   0.64831150  0.2718636
## GI -0.38056872  0.2926510
## P   1.00000000 -0.1670507
## N  -0.16705075  1.0000000
```

From the above residual-index and leverage-index plots, the leverages do not follow a random pattern, since the earlier observations appear to have more leverage. The standardised residuals largely follow a random pattern with the exception of observation 20 which has quite a high standardised residual. Hence (2.1) is likely violated.

Under the normality of errors assumption (2.2), the ordered residuals should be approximately form a straight line with slope 1 and intercept 0 with the quantiles of the standard normal distribution. From examining the Q-Q plot of the standardised residuals against the standard normal distribution, we notice that the sample quantiles mostly appear to match the theoretical quantiles with the exception of the last 3 points, so again it is likely (2.2) is not satisfied.

Assumption (2.3) is assumed to be true by least squares in an intercept model.

The standard residuals should be uncorrelated with the predictor variables/response. As such when plotting the std. residuals against each of the predictors, we expect to see a random scatter of points, which is true for I, D, Response, GI, P and N but NOT for W (although this can likely be forgiven since there are very few points where $W = 1$.) At every fitted value, the spread of the residuals is roughly the same with an outlying point or two in some of the plots. Hence the constant variance assumption (2.4) is satisfied.

Analysis of the std residuals vs predictors/fitted values plot shows no discernible pattern so it is likely that assumption (1) on linearity is satisfied.

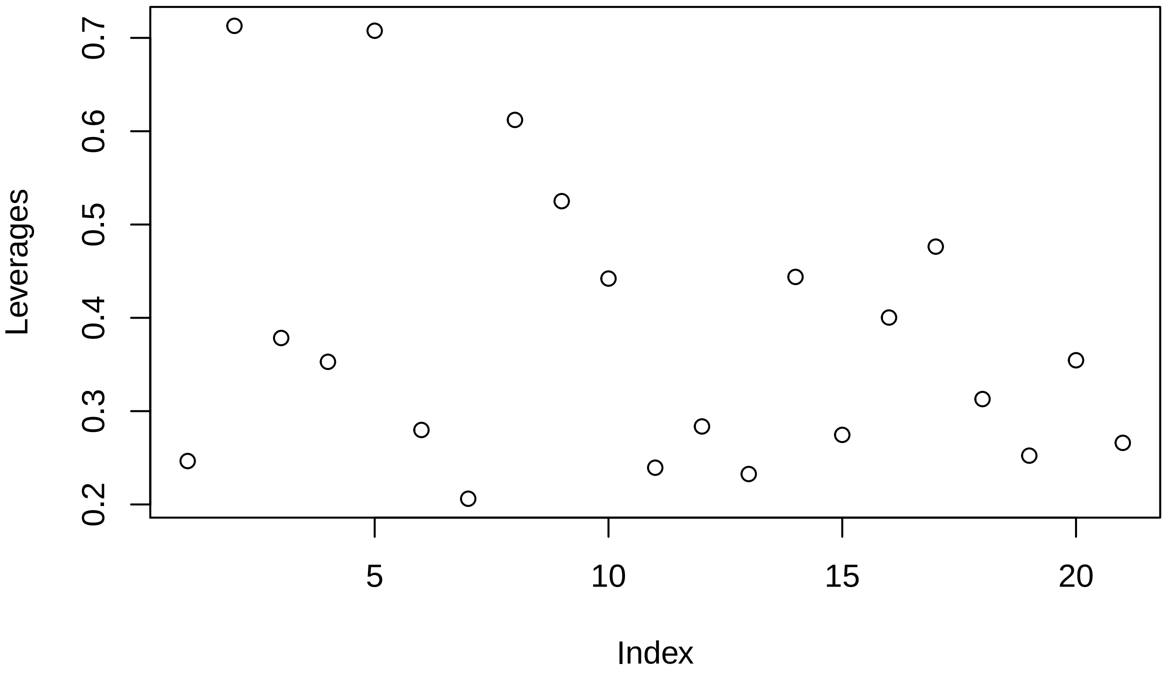
Of these assumptions, (3.2) and (3.1) are difficult to assume. Let us examine the independence of the predictors, i.e. assumption (3.3). Same as with the original model, besides $Cor(D_1, I) = 0.87479$ and $Cor(W, P) = 0.648$, the pairwise correlations between predictors is low. So (3.3) is likely satisfied.

By analyzing the Cook's distance, DFITS and Hadi's plots, we see that observation 20 is more influential than the rest, hence assumption (4) is violated.

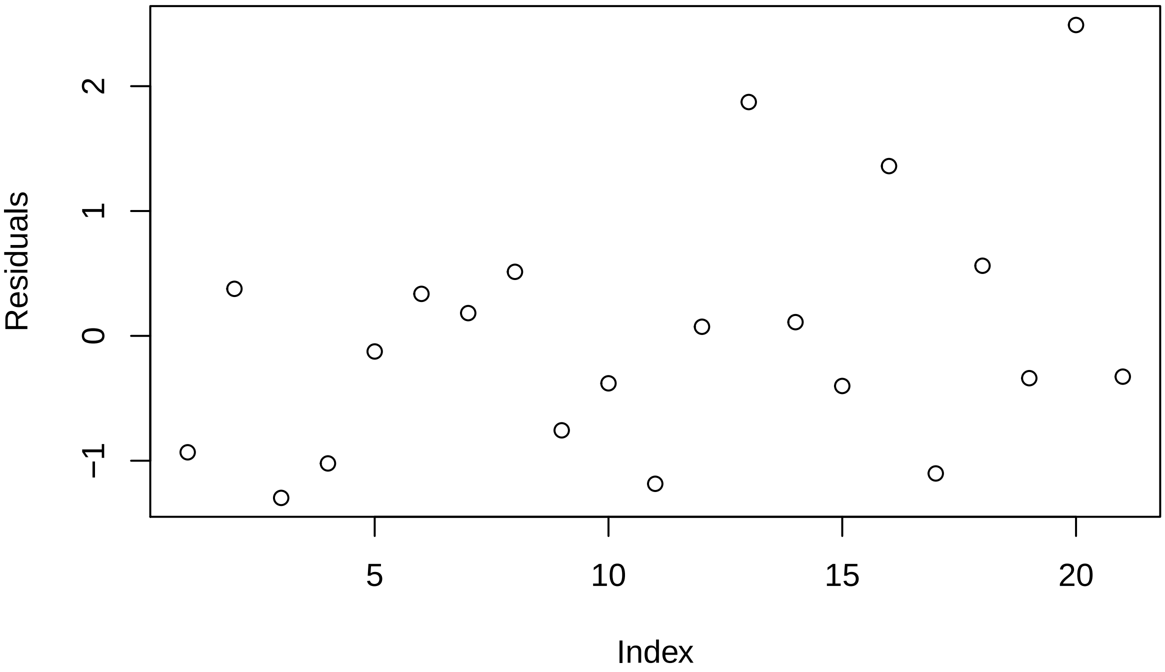
```r
elect_data_q6 <- elect_data_q3a

elect_data_q6$Y <- log((elect_data_q6$V)/(1-elect_data_q6$V)) # by default is natural log
elect_data_q6 <- subset(elect_data_q6,select=c("Y","I","D","W","GI","P","N")) # drop V
elect_model_q6 = lm(Y ~ ., data = elect_data_q6)
summary(elect_model_q6)
```

```
##
## Call:
## lm(formula = Y ~ ., data = elect_data_q6)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.17737 -0.08949 -0.01160  0.04901  0.34337
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.020861   0.147267   0.142 0.889525
## I           -0.084033   0.070707  -1.188 0.255909
## D-1         -0.191349   0.118040  -1.621 0.128999
## D1           0.257358   0.126235   2.039 0.062351 .
## W            0.046266   0.176684   0.262 0.797534
## GI           0.038279   0.007918   4.835 0.000326 ***
## P           -0.002847   0.016714  -0.170 0.867380
## N           -0.020408   0.015911  -1.283 0.222025
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1716 on 13 degrees of freedom
## Multiple R-squared:  0.794,  Adjusted R-squared:  0.6831
## F-statistic: 7.159 on 7 and 13 DF,  p-value: 0.001239
```

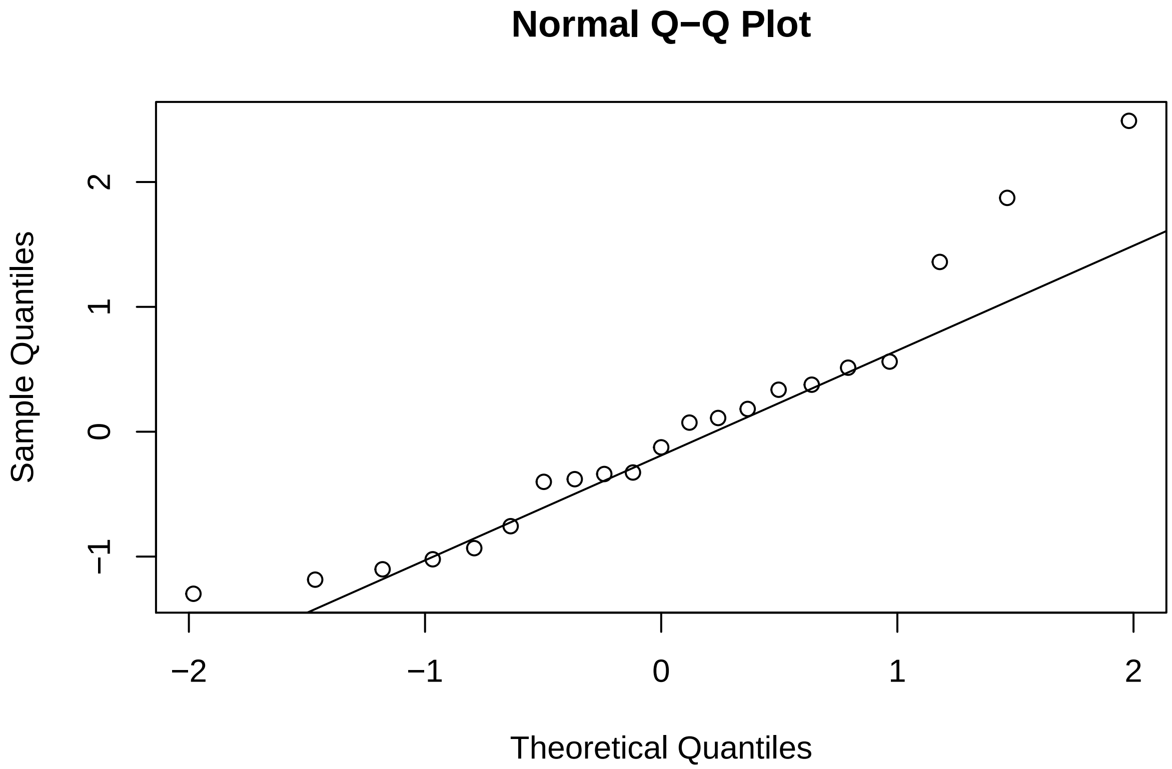
```
leverages <- hatvalues(elect_model_q6) # compute leverages for model Y
elect_model_q6.residuals <- rstandard(elect_model_q6)

plot(seq(1,length(elect_data$V)), leverages, xlab="Index", ylab="Leverages")
```



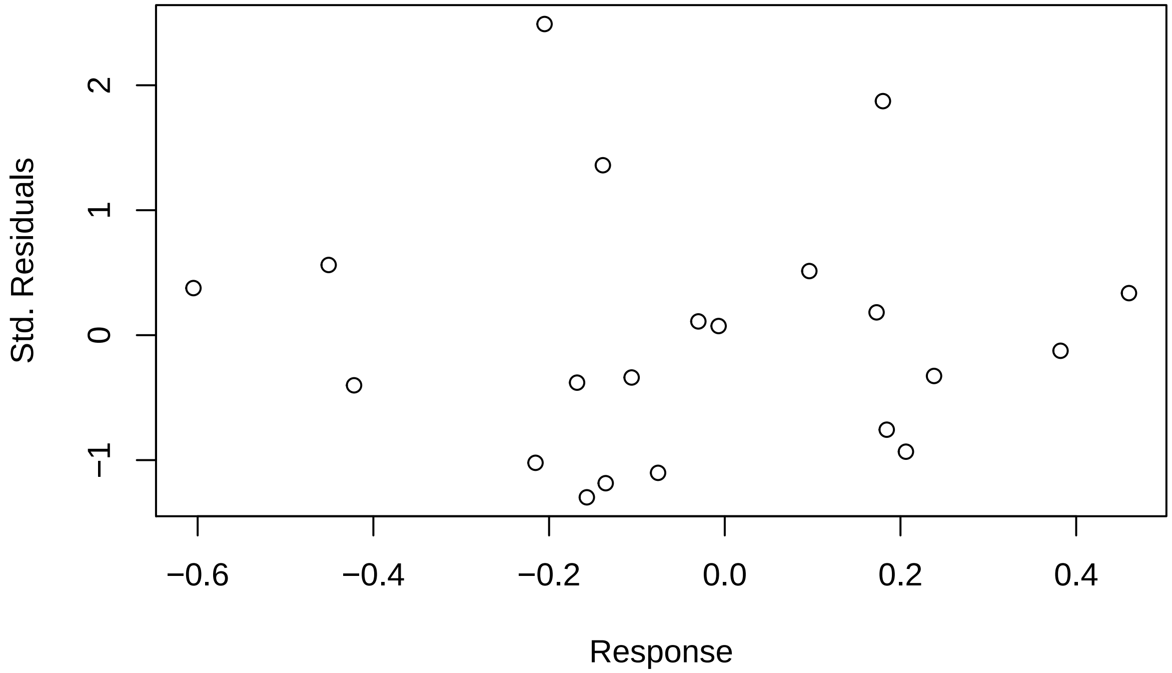
```
plot(seq(1,length(elect_data$V)), elect_model_q6.residuals, xlab="Index", ylab="Residuals")
```



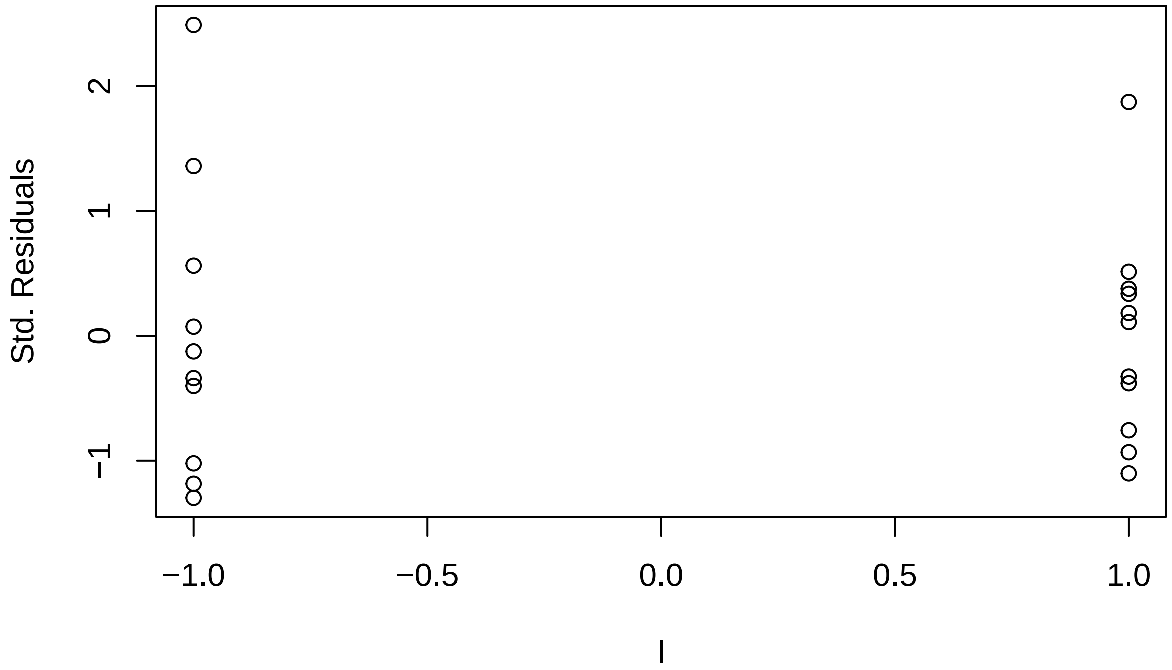
```
qqnorm(elect_model_q6.residuals, ylab="Sample Quantiles", xlab="Theoretical Quantiles")
qqline(elect_model_q6.residuals)
```



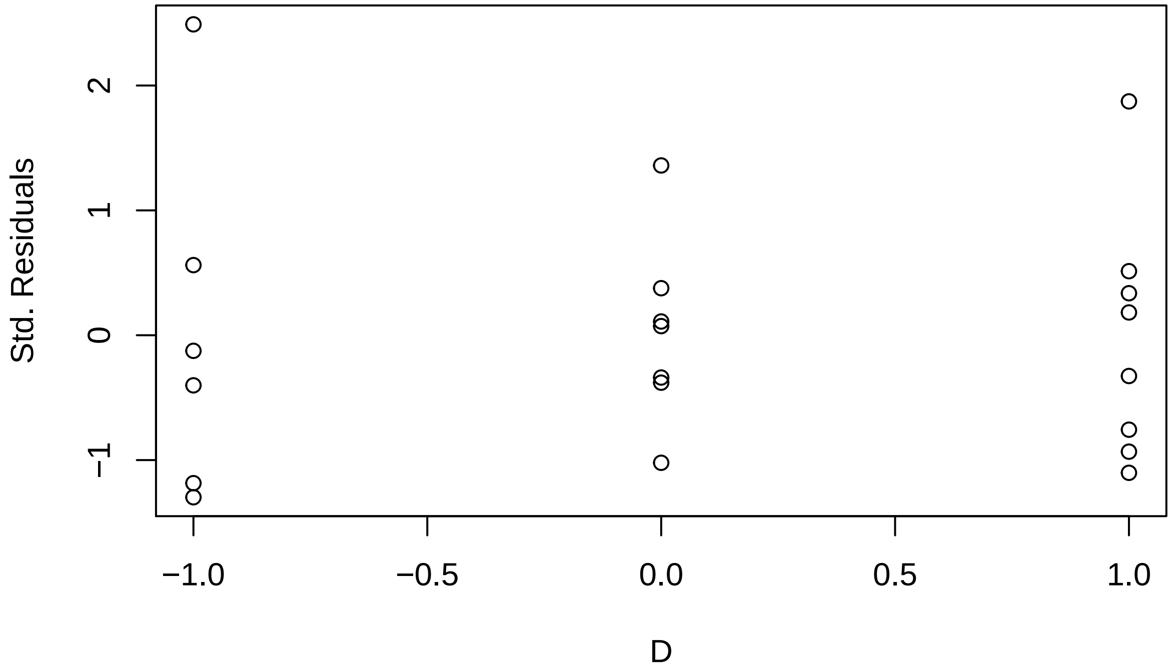
```
plot(fitted.values(elect_model_q6), elect_model_q6.residuals, ylab="Std. Residuals", xlab="Response")
```



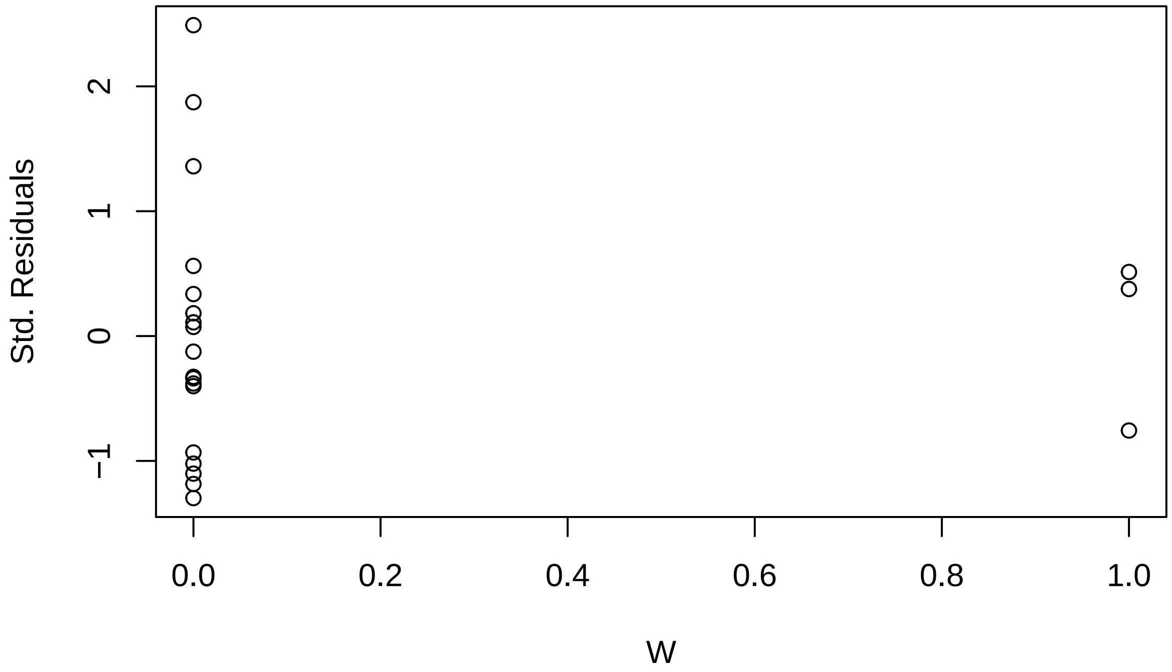
```
plot(elect_data_2a$I, elect_model_q6.residuals, ylab="Std. Residuals", xlab="I")
```



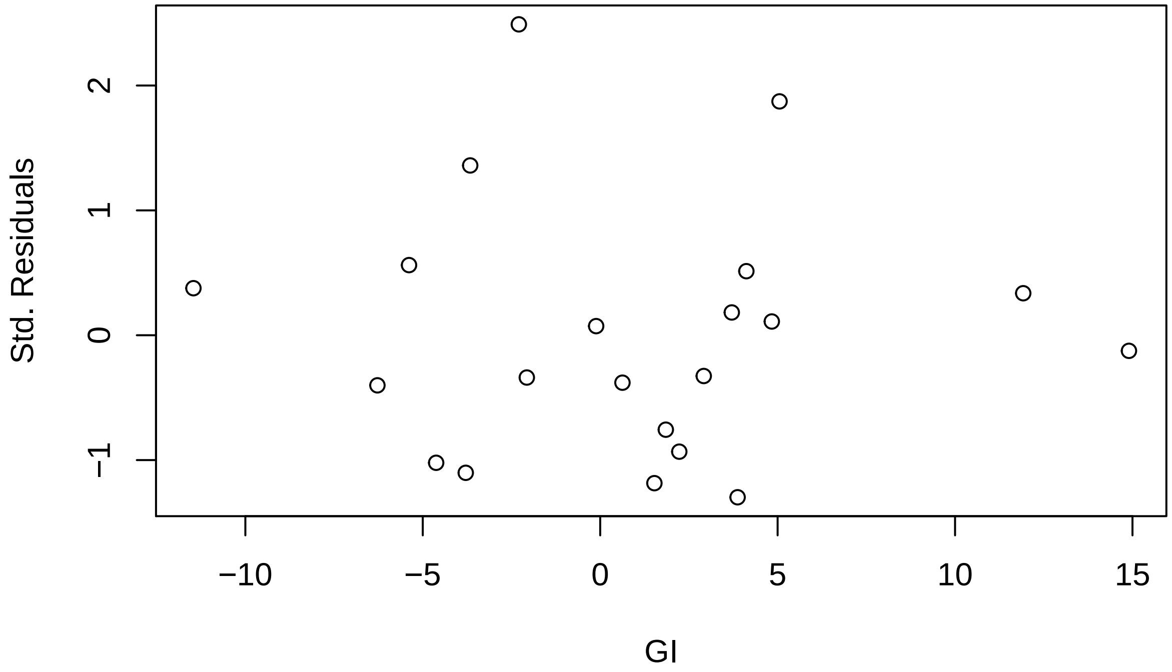
```
plot(elect_data_2a$D, elect_model_q6.residuals, ylab="Std. Residuals", xlab="D")
```



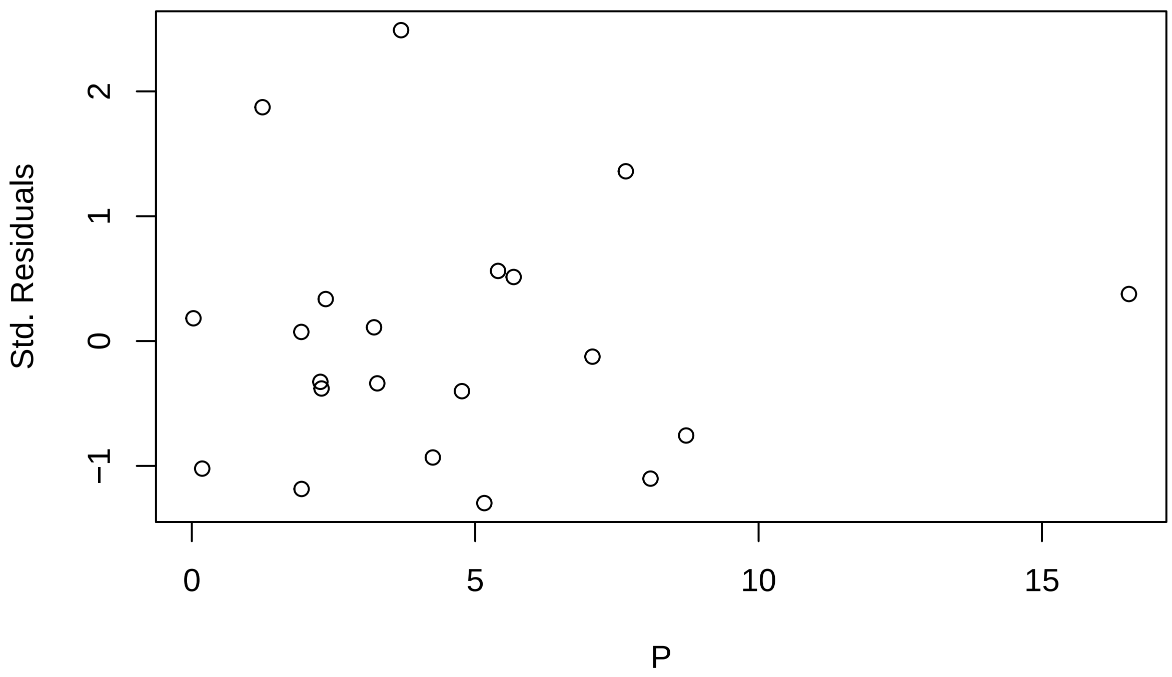
```
plot(elect_data_2a$W, elect_model_q6.residuals, ylab="Std. Residuals", xlab="W")
```



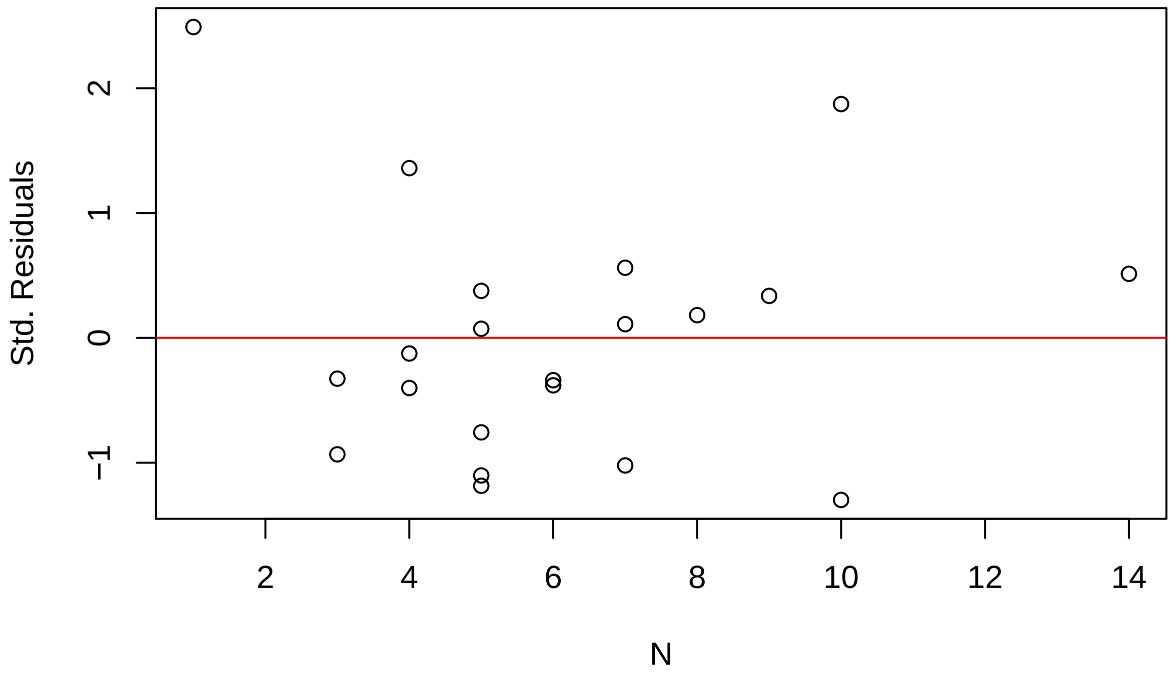
```
plot(elect_data_2a$GI, elect_model_q6.residuals, ylab="Std. Residuals", xlab="GI")
```



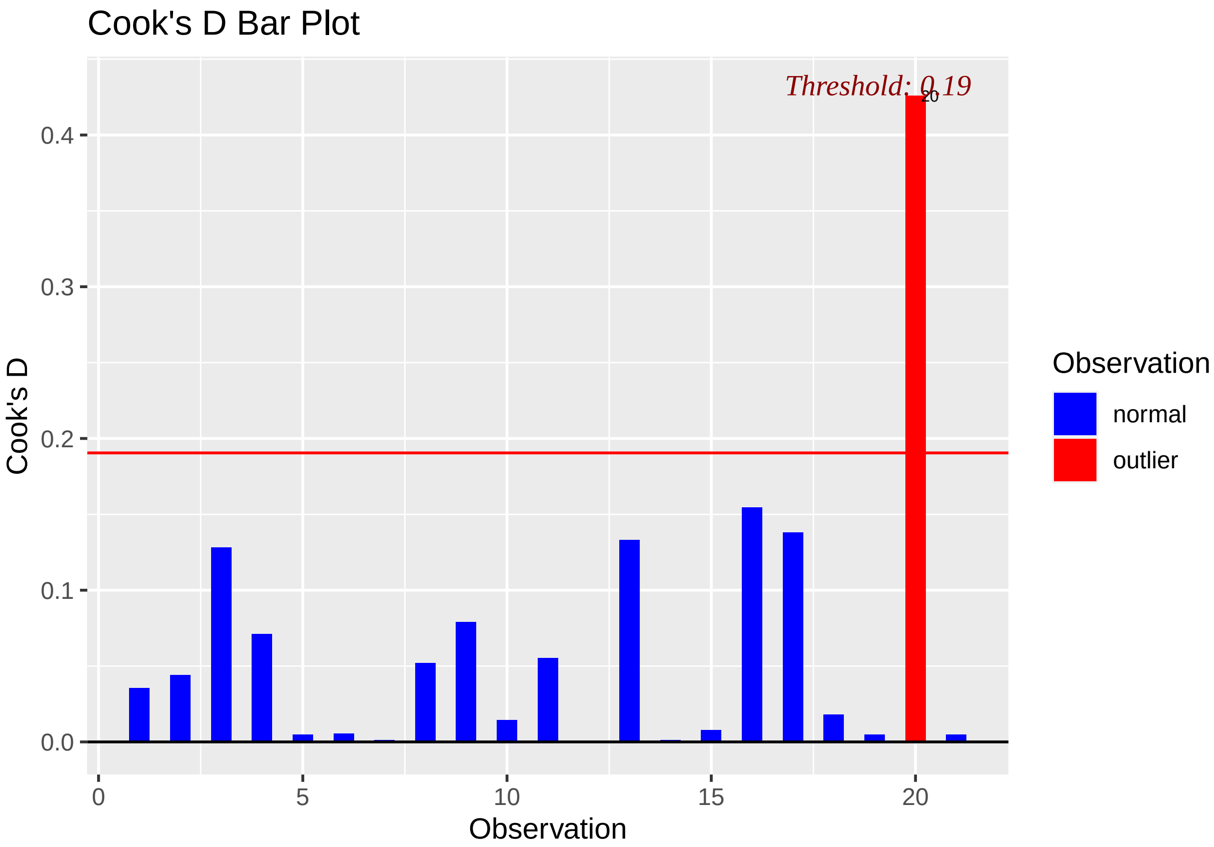
```
plot(elect_data_2a$P, elect_model_q6.residuals, ylab="Std. Residuals", xlab="P")
```



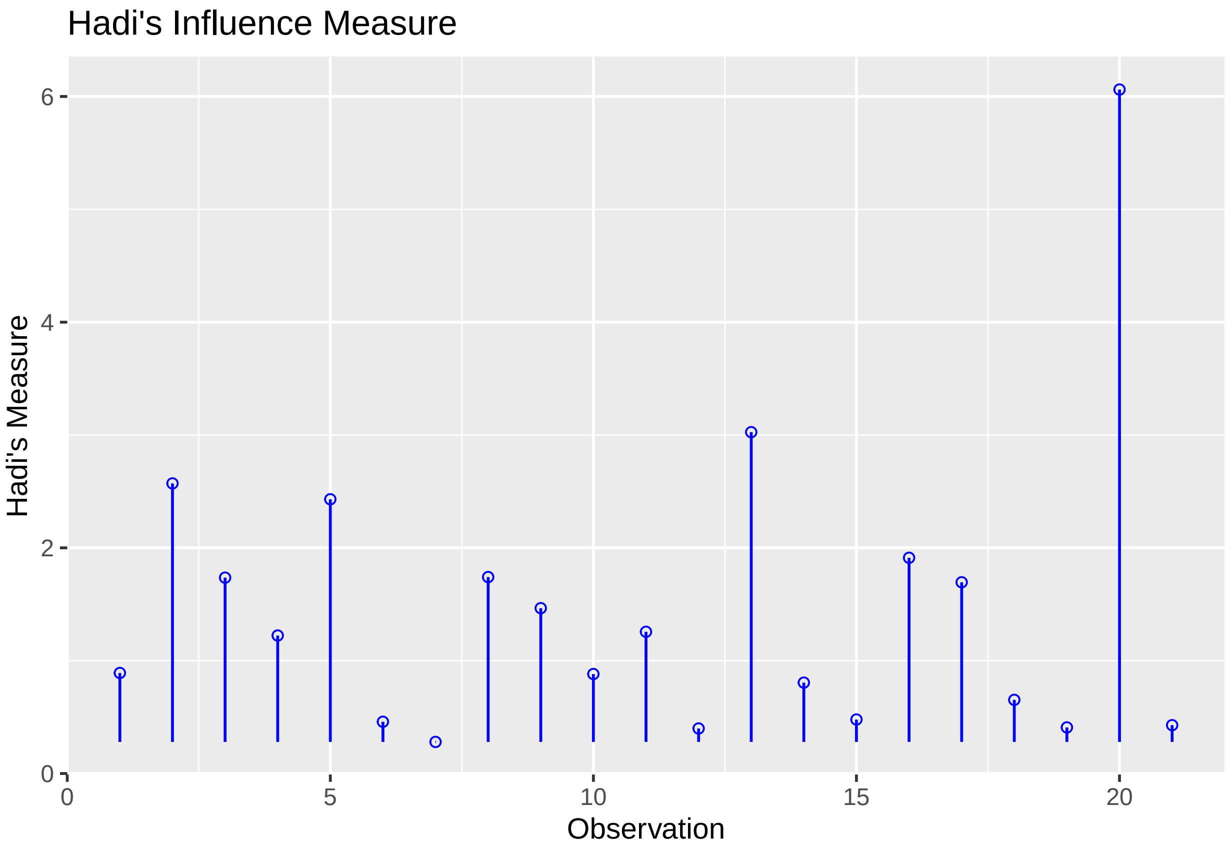
```
plot(elect_data_2a$N, elect_model_q6.residuals, ylab="Std. Residuals", xlab="N")
abline(a=0, b=0, col="red")
```



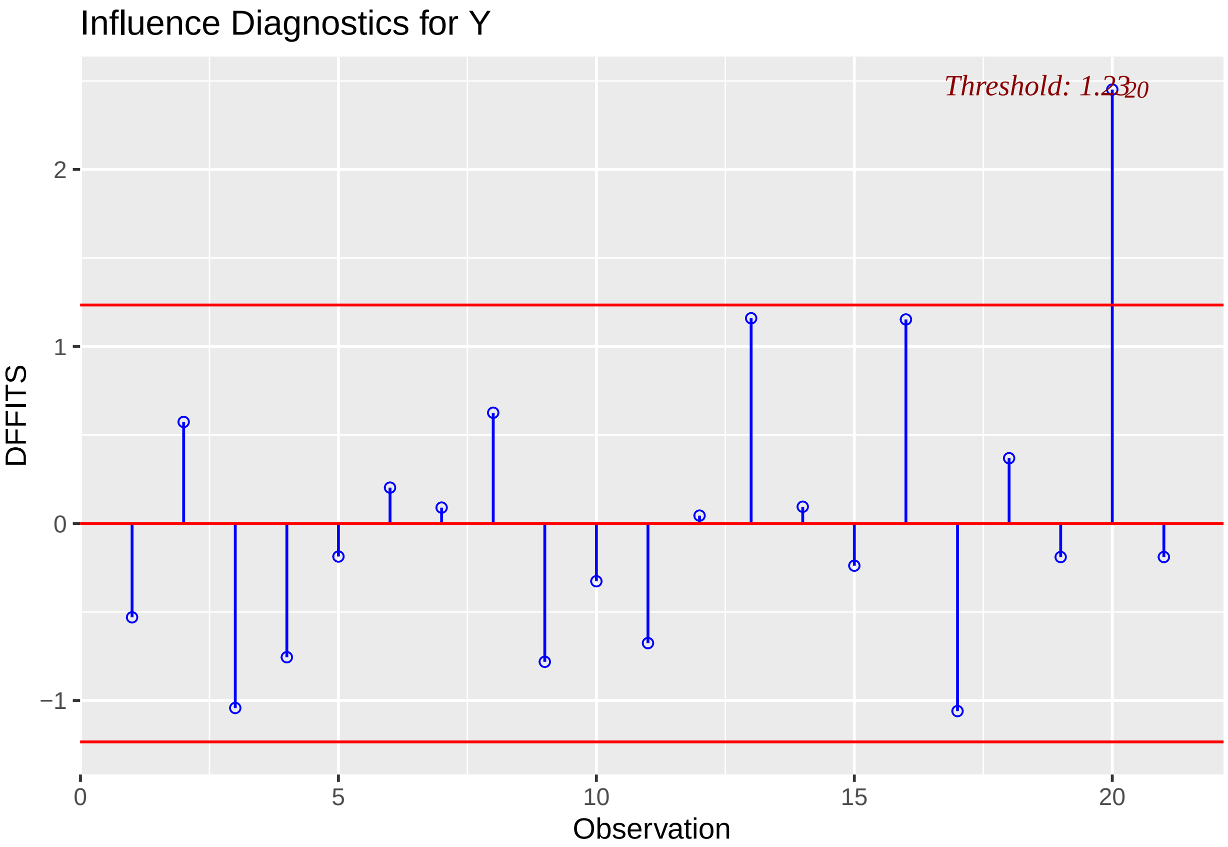
```
ols_plot_cooksd_bar(elect_model_q6)
```



```
ols_plot_hadi(elect_model_q6)
```



```
ols_plot_dffits(elect_model_q6)
```



From the above residual-index and leverage-index plots, the leverages do not follow a random pattern, since the earlier observations appear to have more leverage. The standardised residuals largely follow a random pattern with the exception of observation 20 which has quite a high standardised residual. Hence (2.1) is likely violated.

Under the normality of errors assumption (2.2), the ordered residuals should be approximately form a straight line with slope 1 and intercept 0 with the quantiles of the standard normal distribution. From examining the Q-Q plot of the standardised residuals against the standard normal distribution, we notice that the sample quantiles mostly appear to match the theoretical quantiles with the exception of the last 3 points, so again it is likely (2.2) is not satisfied.

Assumption (2.3) is assumed to be true by least squares in an intercept model.

The standard residuals should be uncorrelated with the predictor variables/response. As such when plotting the std. residuals against each of the predictors, we expect to see a random scatter of points, which is true for I, D, Response, GI, P and N but NOT for W (although this can likely be forgiven since there are very few points where $W = 1$.) At every fitted value, the spread of the residuals is roughly the same with an outlying point or two in some of the plots. Hence the constant variance assumption (2.4) is satisfied.

Analysis of the std residuals vs predictors/fitted values plot shows no discernible pattern so it is likely that assumption (1) on linearity is satisfied.

Of these assumptions, (3.2) and (3.1) are difficult to assume. Let us examine the independence of the predictors, i.e. assumption (3.3). Same as with the original model, besides $Cor(D_1, I) = 0.87479$ and $Cor(W, P) = 0.648$, the pairwise correlations between predictors is low. So (3.3) is likely satisfied.

By analyzing the Cook's distance, DFITS and Hadi's plots, we see that observation 20 is more influential than the rest, hence assumption (4) is violated.

There is very little difference in the plots for Q6 between using V as opposed to using Y.

However, when comparing the plots with those in Q5, we see the DFITS plots in Q6 has more of the points with $DFITS > 1$ (influence measure) namely 3, 13, 16, and 17. However, when compared with the DFITS plot in Q5 has slightly lower DFITS for points 3, 16 and 17, but a higher influence for observation 13. If we ignore point 20 in the Q5 and Q6 Hadi's plots, the spread of influence values is overall lower and more balanced for the Hadi's plots in Q6. The rest of the plots (stdres-index, leverage-index, QQ plot of stdres vs stdnorm, stdres vs predictors, stdres vs fitted values) seem to be largely the same however.

Q6b) The Q6 analogue of Q5b: The form of the function $V = f(\beta_0 + \beta_1 I + \alpha_1 D_1 + \alpha_2 D_2 + \beta_3 W + \beta_4 (G \cdot I) + \beta_5 P + \beta_6 N + \epsilon)$ is also the logistic function $V = f(Y) = \frac{e^Y}{1 + e^Y}$ where $Y = \beta_0 + \beta_1 I + \alpha_1 D_1 + \alpha_2 D_2 + \beta_3 W + \beta_4 (G \cdot I) + \beta_5 P + \beta_6 N + \epsilon$.