Leung Ko Tsun MATH 3424 HW3

Problem 1.

a. We can see from the ANOVA table, there are $48+1+1=50$ workers in the dataset.

b. $Var(Y) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2 = \frac{SST}{n-1} = \frac{SSR+SSE}{n-1} = \frac{98.8313+338.449}{49} = 8.9241$

c. $\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1\bar{X} = 15.58 - 2.81(0.52) = 14.1188$

d. $50 - 0.52 \times 50 = 50 - 26 = 24$

e. Percentage of variability in Y accounted for by X

$$= \frac{SSR}{SST} \times 100\%$$

$$= \frac{SSR}{SSR+SSE} \times 100\%$$

$$= \frac{98.8313}{98.8313+338.449} \times 100\%$$

$$= 22.60136\%$$

(∵ coefficient for X is -ve)

f). $Cor(X,Y) = -\sqrt{R^2} = -\sqrt{\frac{SSR}{SST}} = -\sqrt{\frac{98.8313}{98.8313+338.449}} = -0.47541$

g). If the worker is a male, his weekly wage in the company is estimated to be \$281 less than that of females.

h. Estimated weekly wages of a men:

$= (15.58 - 2.81 \times 1) \times 100$

$= \$1277.$

i. Estimated weekly wages of a woman:

$= (15.58 - 2.81 \times 0) \times 100$

$= \$1558$

j. From R, $t_{(n-2, \frac{\alpha}{2})} = t_{(48, 0.025)} = 2.010635$

⇒ 95% confidence interval of β_1:

$= [\hat{\beta}_1 - t_{(n-2, \frac{\alpha}{2})} se(\hat{\beta}_1), \hat{\beta}_1 + t_{(n-2, \frac{\alpha}{2})} se(\hat{\beta}_1)]$

$= [-2.81 - 2.010635 \times 0.715, -2.81 + 2.010635 \times 0.715]$

$= [-4.31798, -1.30202]$

k. H_0: $\beta_1 = 0$ (Average weekly wages for men is equal to that of the women)

H_1: $\beta_1 \neq 0$ (Average weekly wages for women is not equal to that of the men)

From R, $t_{(n-2, \frac{\alpha}{2})} = t_{(48, 0.025)} = 2.010635$

From the regression output, t-statistics: $t_1 = -3.74$

$|t_1| = 3.74 > 2.010635$

Therefore, at a significant level of $\alpha = 0.05$, we reject the null hypothesis, i.e. β_1 is statistically different from zero.

⇒ Average weekly wages of men is significantly different from that of the women at a significant level of $\alpha = 0.05$.

Q2

a). When D=-1, the general regression model is:

$$V = \beta_0 + \beta_1 I + \beta_2 \times (1) + \beta_3 W + \beta_4 (G \cdot Z) + \beta_5 P + \beta_6 N + \varepsilon$$

~~From the given data, I, D, W are linearly dependent, as shown below:~~
~~so they don't need to be included in the fitted regression model.~~

~~When D=-1, $V = 0.491548 + 0.000773 (G \cdot Z) + 0.000854 (1) - 0.010574 (N)$~~

When D=0, the general regression model is:

$$V = \beta_0 + \beta_1 I + \beta_2 \times (0) + \beta_3 W + \beta_4 (G \cdot Z) + \beta_5 P + \beta_6 N + \varepsilon$$

~~From the given data, the fitted regression model is:~~

$V =$ ~~0.531118 - 0.023745 I - 0.044262 + 0.11999 (G·Z) + 0.000650 P - 0.013391 N~~

When D=1, the general regression model is:

$$V = \beta_0 + \beta_1 I + \beta_2 \times 1 + \beta_3 W + \beta_4 (G \cdot Z) + \beta_5 P + \beta_6 N + \varepsilon$$

~~From the given data, I and D are linearly dependent, as shown below:~~
~~so they don't need to be included in the fitted regression model.~~

~~When D=-1, $V = 0.529041 + 0.029520 W + 0.009378 (G \cdot Z) - 0.000676 (P) + 0.004448 (N)$~~

Using full data, the fitted regression model is:

$$V = 0.511162 - 0.020107 I + 0.054615 D + 0.013395 W + 0.009690 (G \cdot Z) - 0.000724 P - 0.005182 N$$

Q2(cont'd).

Interpretation for β_2:

Default case (when D=0), there is no democratic or republic incumbent running for election.

If the Democratic incumbent is running, the estimated Democratic share (V) will increase by β_2 (0.0546159), while keeping other predictors as constant.

If the Republic incumbent is running, then the estimated Republic share (V) will decrease by β_2 (0.0546159), while keeping other predictors as constant.

Q2b). No.
From the model for full data computed in (a), the p-value of I is 0.2539, which is > 0.05, so it is insignificant.
Also, I & D are very similar in nature, so just include D is enough.

Q2c). Yes.
From the model for full data computed in (b), the p-value of (G·Z) is 8.24e-4, which is < 0.05, so it is significant.
It should be included in the regression model.

Q2d). We can do the following things to achieve for the best model:
1. Remove one of D & I, as explained in Q2(b). (I removed I)
2. Remove W, as there is not enough data to quantify its effect, only having 3 observations with W=1.
3. P·D, N·D, G·D should be included, since good economic performance should not help a Democratic candidate if the incumbent is Republic. They should be considered together.

Q2a R result:

```
Call:
lm(formula = V ~ I + D + W + GI + P + N, data = pres_data)

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
```

(Q2d). Cont'd). So, our alternative model is:

$$V = \beta_0 + \beta_1 \cdot D + \beta_2 \cdot G + \beta_3 (G \cdot D) + \beta_4 \cdot P + \beta_5 (P \cdot D) + \beta_6 \cdot N + \beta_7 (N \cdot D) + \varepsilon.$$

Fitting this model. we have:

$$\hat{V} = 0.52594 - 0.007288D + 0.008515G + 0.011224(G \cdot D) - 0.002191P - 0.003299(P \cdot D) - 0.008774N + 0.008089(N \cdot D)$$

From result of R, only G·D, N, and N·D are statistically significant at $\alpha = 0.05$.

attached below

Then, we try to fit the model as: $\hat{V} = \beta_0 + \beta_1 \cdot G + \beta_2 \cdot (G \cdot D) + \beta_3 N + \beta_4 (N \cdot D)$, we have:

attached below

$$\hat{V} = 0.51947 + 0.004099G + 0.011379(G \cdot D) - 0.008288 \cdot N + 0.005384(N \cdot D)$$

then all coefficients are statistically significant, and F-test indicates that this reduction is acceptable.

Interpretation of the terms:

G·D, N·D indicates whether a candidate will benefit by satisfactory leadership in economics by the candidate's party.

G·N indicates the difference of the level to which a candidate can benefit from such leadership for Democratic or Republic.

```
Call:
lm(formula = V ~ D + G + GD + P + PD + N + ND, data = pres_data)

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
G            0.002815   0.002272   1.239 0.237245
GD           0.010224   0.002386   4.285 0.000887 ***
P           -0.002191   0.003215  -0.682 0.507493
PD          -0.002299   0.005010  -0.459 0.653957
N           -0.008774   0.003435  -2.554 0.023989 *
ND           0.008089   0.003465   2.335 0.036243 *
---
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
```

)2d(2)

```
> model_q2_d2 <- lm(V~G+GD+N+ND, data = pres_data)
> summary(model_q2_d2)

Call:
lm(formula = V ~ G + GD + N + ND, data = pres_data)

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 ***
G            0.004099   0.001641   2.498   0.0238 *
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 **
---
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
```

Q3 a).

The general regression model is:

$$V = \beta_0 + \beta_1 I + \alpha_1 D_1 + \alpha_2 D_2 + \beta_3 W + \beta_4 (G \cdot I) + \beta_5 P + \beta_6 N + \varepsilon$$

The fitted regression model is shown as follows:

When $D=1$,

$$\hat{V} = 0.5054760 - 0.0205982 I + 0.0633485 D_1 + 0.0123948 W + 0.0009422 (G \cdot I) - 0.0006963 P - 0.0051083 N$$

When $D=-1$,

$$\hat{V} = 0.5054760 - 0.0205982 I - 0.0469714 D_2 + 0.0123948 W + 0.0009422 (G \cdot I) - 0.0006963 P - 0.0051083 N$$

When $D=0$,

$$\hat{V} = 0.5054760 - 0.0205982 I + 0.0123948 W + 0.0009422 (G \cdot I) - 0.0006963 P - 0.0051083 N$$

Intrepretation for α_1 and α_2:

Default case (when $D=0$), there is no democratic or republic incumbent running for selection.

When $D=1$, if a Democratic incumbent is running for election, V will increase by α_1 (0.0633485) unit, while keeping other predictors as constant.

When $D=-1$, if a Republic incumbent is running for election, V will decrease by α_2 (-0.0469714) unit, while keeping other predictors as constant.

13 R result

```
> model_q3_full <- lm(V~I+D1+D2+W+GI+P+N, data = pres_data)
> summary(model_q3_full)

Call:
lm(formula = V ~ I + D1 + D2 + W + GI + P + N, data = pres_data)

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           0.0633485  0.0312177   2.029 0.063423 .
D2          -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
```

Q3(c) R Result

```
> pres_data$D1_minus_D2 <- pres_data$D1 - pres_data$D2
> model_q3_reduced <- lm(V~I+D1_minus_D2+W+GI+P+N, data = pres_data)
> anova(model_q3_reduced, model_q3_full)
Analysis of Variance Table

Model 1: V ~ I + D1_minus_D2 + W + GI + P + N
Model 2: V ~ I + D1 + D2 + W + GI + P + N
  Res.Df      RSS Df  Sum of Sq      F Pr(>F)
1     14 0.023686
2     13 0.023423  1 0.00026227 0.1456  0.709
```

Q3b). As we have $\alpha_1 = -\alpha_2$, the model becomes:

$$V = \beta_0 + \beta_1 I - \alpha_2 D_1 + \alpha_2 D_2 + \beta_3 W + \beta_4 (G \cdot I) + \beta_5 P + \beta_6 N + \varepsilon$$

$$V = \beta_0 + \beta_1 I - \alpha_2 (D_1 - D_2) + \beta_3 W + \beta_4 (G \cdot I) + \beta_5 P + \beta_6 N + \varepsilon$$

Consider 3 values of D,

$$\begin{cases} \text{when } D=1,\ D_1 - D_2 = 1 - 0 = 1 \\ \text{when } D=0,\ D_1 - D_2 = 0 - 0 = 0 \\ \text{when } D=-1,\ D_1 - D_2 = 0 - 1 = -1 \end{cases}$$

c) We can observe that the behaviours of $D_1 - D_2$ is same as D, so for all 3 situations, the model in Problem 2 can be obtained as a special case in Problem 3 by assuming $\alpha_1 = -\alpha_2$.

c. $\begin{cases} H_0 : \alpha_1 = -\alpha_2 \\ H_1 : \alpha_1 \neq -\alpha_2 \end{cases}$

From the result of R:

F value is 0.1456, which is less than $F(1,13;0.05) = 4.667$, so we cannot reject H_0.

⇒ the given dataset supports the assumption that $\alpha_1 = -\alpha_2$.

Q4 a). From the table, $Corr(Y, X) = Corr(Y', X) = -0.777$

b). From the table, when λ drops from 1 to −1, correlation coefficient increases from −0.777 to 0.999. Then, the correlation coefficient drops from 0.999 to 0.943 when λ drops from −1 to −2. Since the most important purpose of transformation is to achieve linearity → strong correlation, we can assert that best λ is −1.

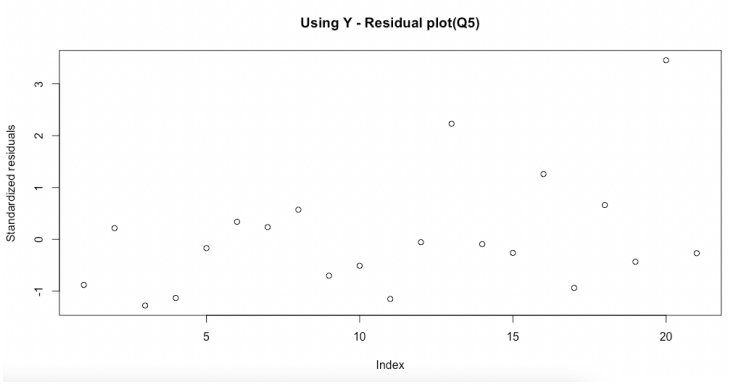
c. $\frac{1}{Y} = \beta_0 + \beta_1 X + \varepsilon$

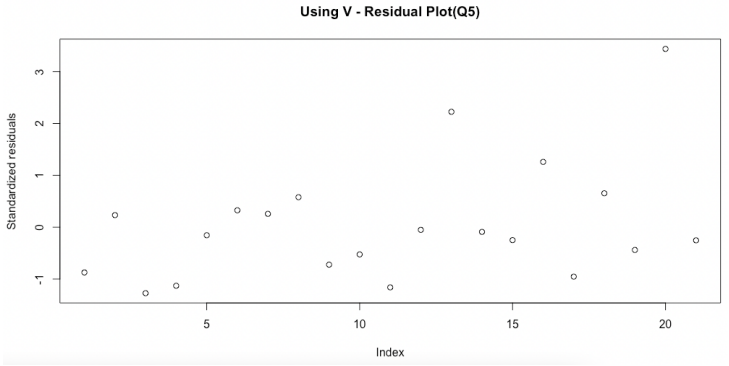
25).

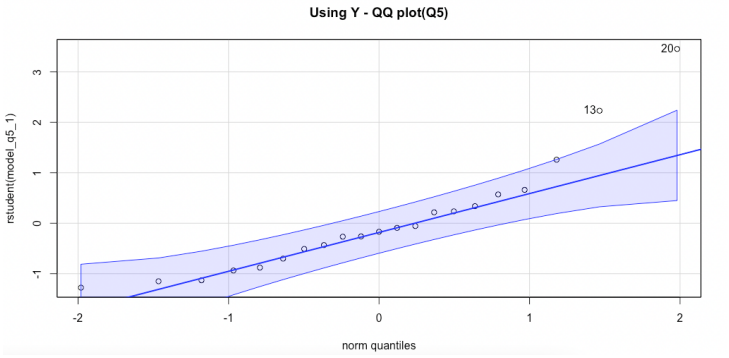
Regression result:

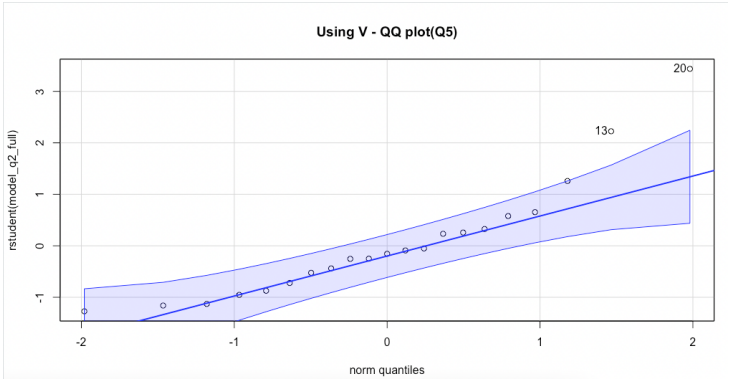
```
> #q5
> pres_data$Y <- log(pres_data$V / (1-pres_data$V))
> model_q5_1 <- lm(Y~I+D+W+GI+P+N, data = pres_data)
> summary(model_q5_1)

Call:
lm(formula = Y ~ I + D + W + GI + P + N, data = pres_data)

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
```

The plot of standardized residuals by index and Q-Q plot is shown on the next page:









The plots for models using V and Y are similar. And we can conclude that two models satisfy the standard assumptions to the same extent.

$$Y = \ln\left(\frac{V}{1-V}\right)$$

$$e^{Y} = \frac{V}{1-V}$$

$$V = \frac{e^{Y}}{1+e^{Y}}$$

$$V = \frac{e^{\beta_0 + \beta_1 I + \beta_2 D + \beta_3 W + \beta_4 (G \cdot I) + \beta_5 P + \beta_6 N + \varepsilon}}{1 + e^{\beta_0 + \beta_1 I + \beta_2 D + \beta_3 W + \beta_4 (G \cdot I) + \beta_5 P + \beta_6 N + \varepsilon}}$$

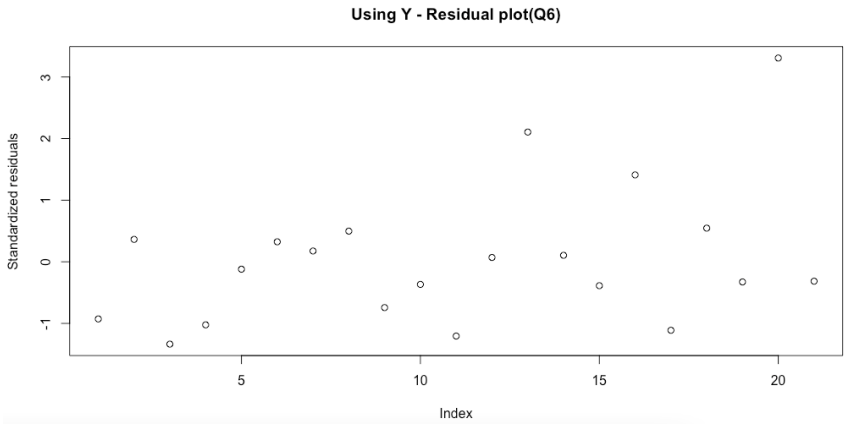
i.e. $V = f(\beta_0 + \beta_1 I + \beta_2 D + \beta_3 W + \beta_4 (G \cdot I) + \beta_5 P + \beta_6 N + \varepsilon)$,,

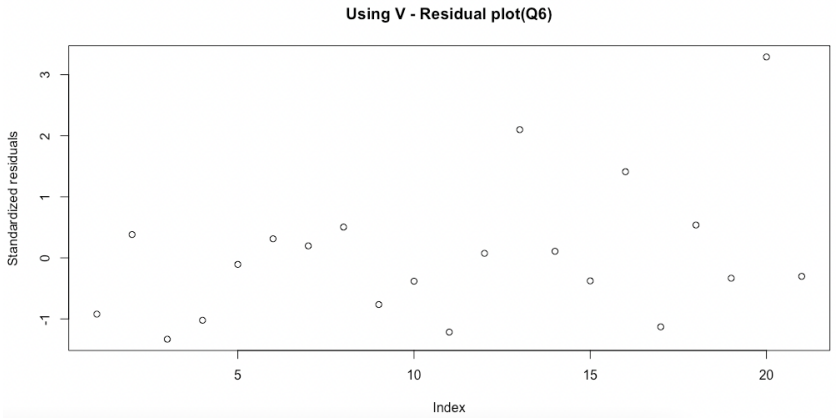
Q6.

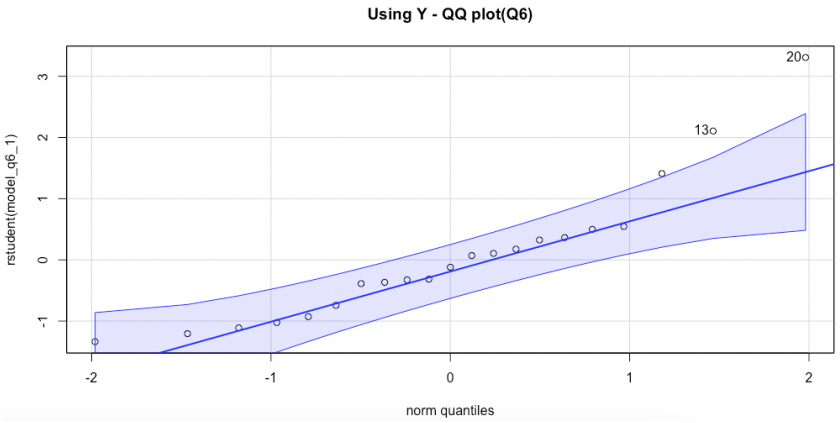
Regression Result:

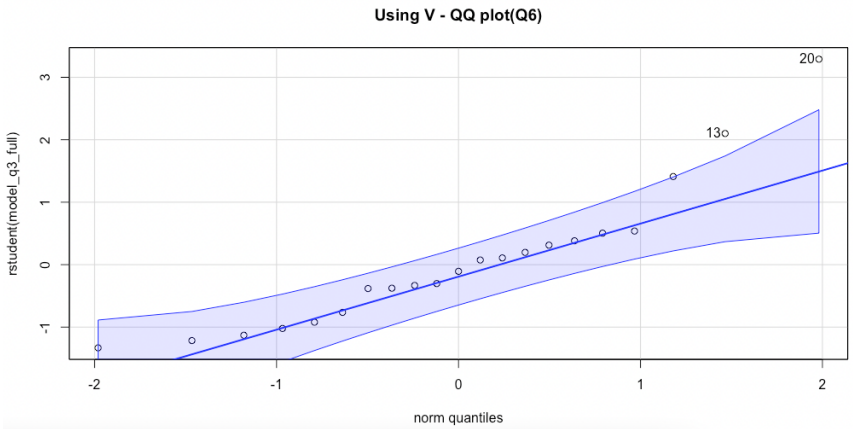
```
> model_q6_1 <- lm(Y~I+D1+D2+W+GI+P+N, data = pres_data)
> summary(model_q6_1)

Call:
lm(formula = Y ~ I + D1 + D2 + W + GI + P + N, data = pres_data)

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
D1           0.257358   0.126235   2.039 0.062351 .
D2          -0.191349   0.118040  -1.621 0.128999
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
```

The plot of standardized residuals by index and Q-Q plot are shown below:









The plots are similar for both models. We can conclude that two models satisfy the standard assumptions to the same extent.