

MATH 588

HW10

Md Ismail Hossain

4/27/2022

Question 1

```
library(Sleuth3)
head(ex2115)
```

##	Context	Mode	Level	Number	InFavor	NotInFavor	PercentInFavor
## 1	Vietnam	scattered	high	132	27	105	20.5
## 2	Cuba	scattered	high	132	41	91	31.1
## 3	Vietnam	scattered	low	132	37	95	28.0
## 4	Cuba	scattered	low	131	51	80	38.9
## 5	Vietnam	not	high	132	45	87	34.1
## 6	Cuba	not	high	131	64	67	48.9

```
attach(ex2115)
```

a

```
ex2115$prop <- PercentInFavor/100
binResponse <- cbind(InFavor, NotInFavor)
glm1_1 <- glm (binResponse ~ as.factor(Level), data = ex2115, family = binomial)
summary (glm1_1)
```

```
##
## Call:
## glm(formula = binResponse ~ as.factor(Level), family = binomial,
##      data = ex2115)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -3.3332  -1.7098  -0.2479   1.5971   3.6001
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.681783    0.092233  -7.392 1.45e-13 ***
## as.factor(Level)low  0.008839    0.130907   0.068  0.946
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 41.915  on 7  degrees of freedom
## Residual deviance: 41.911  on 6  degrees of freedom
## AIC: 87.243
##
## Number of Fisher Scoring iterations: 4
```

The p-value for Level is not significant at 5% level of significance. So , we can conclude that proportion of favorable response to the target question does not depend on the wording (LEVEL) of the question.

b

```
glm1_2 <- glm (binResponse ~ as.factor(Context), data = ex2115, family = binomial)
summary (glm1_2)
```

```
##
```

```
## Call:
## glm(formula = binResponse ~ as.factor(Context), family = binomial,
##      data = ex2115)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.36138  -1.00456  -0.04602   1.21437   1.92005
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.36373    0.08942  -4.068 4.75e-05 ***
## as.factor(Context)Vietnam -0.65564    0.13310  -4.926 8.40e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 41.915  on 7  degrees of freedom
## Residual deviance: 17.234  on 6  degrees of freedom
## AIC: 62.567
##
## Number of Fisher Scoring iterations: 4
```

Here, the p-value for both the intercept and factor variable are statistically significant because the p-value is less than 0.05. The odds of beta 1 is 0.519. So, the odds of the context question preceding the target question about U.S. aid to the Nicaraguan Contra rebels of Cuba is almost 2 times than the Vietnam.

c

```
ind_mode <- ifelse (Mode == "scattered", 1, 0)
ind_context <- ifelse (Context == "Vietnam", 1, 0)
glm1_3 <- glm (binResponse ~ ind_context + ind_mode, data = ex2115, family = binomial)
summary (glm1_3)
```

```
##
## Call:
## glm(formula = binResponse ~ ind_context + ind_mode, family = binomial,
##      data = ex2115)
##
## Deviance Residuals:
##      1       2       3       4       5       6       7       8
## -0.64635  -1.30050   1.40199   0.58876   0.95115   0.71714  -1.73233  -0.04465
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.1713    0.1112  -1.541  0.12341
## ind_context   -0.6644    0.1337  -4.968 6.75e-07 ***
## ind_mode      -0.3849    0.1333  -2.888  0.00388 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 41.9151  on 7  degrees of freedom
```

```
## Residual deviance:  8.8432  on 5  degrees of freedom
## AIC: 56.176
##
## Number of Fisher Scoring iterations: 4
```

Here, the two-sided p-value for the β_1 and β_2 are $6.75e-0$ and 0.00388 respectively which are significant at 5% level of significance. SO, we can conclude that the proportion depend on the context question to different extents according to whether the target and context questions are scattered.

```
attach (ex2116)
prop_tumor <- Tumor / Total
odds_tumor <- prop_tumor / (1 - prop_tumor)
odds_tumor
```

```
## [1] 0.11538462 0.06172840 0.02298851 0.11842105 0.53571429 0.91111111
## [7] 0.45762712 0.62962963 1.54285714 1.60606061 2.46153846 1.66666667
## [13] 4.17647059 4.56250000 2.82608696 4.00000000 3.30000000 10.71428571
## [19] 8.00000000 4.56250000
```

Exercise 21.16

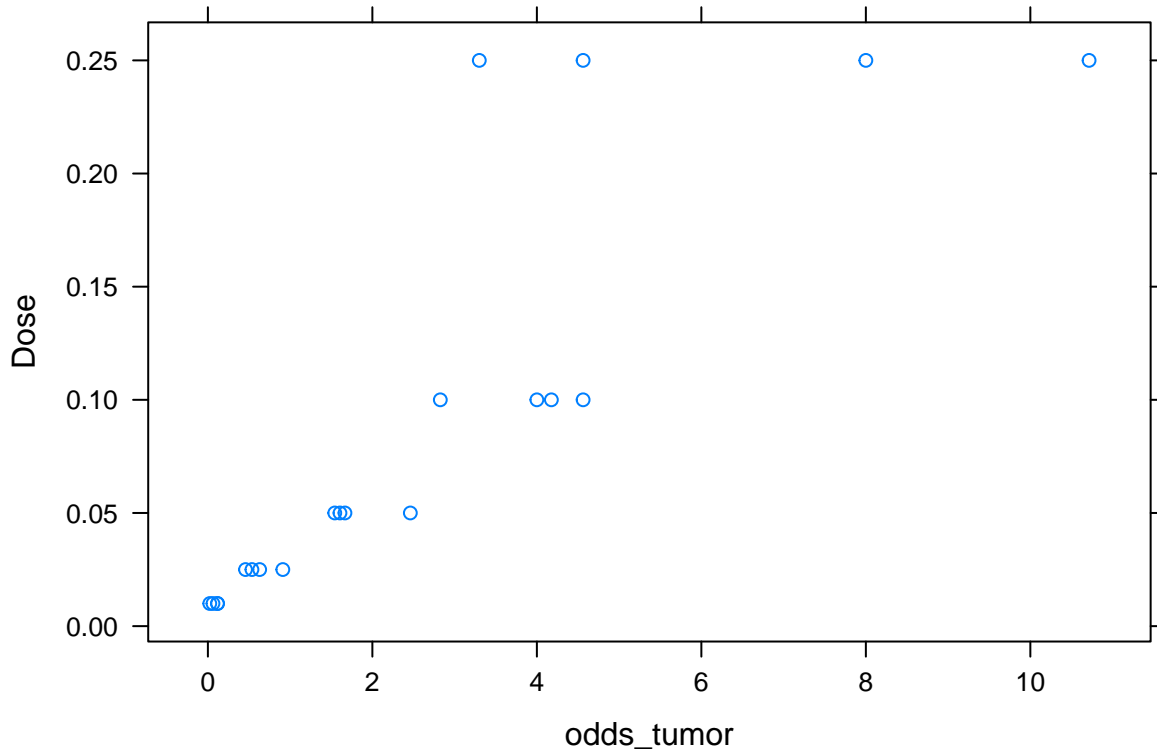
```
head (ex2116)
```

```
##      Dose Tumor Total
## 1 0.010      9    87
## 2 0.010      5    86
## 3 0.010      2    89
## 4 0.010      9    85
## 5 0.025     30    86
## 6 0.025     41    86
```

```
attach (ex2116)
prop_tumor <- Tumor / Total
odds_tumor <- prop_tumor / (1 - prop_tumor)
odds_tumor
```

```
## [1] 0.11538462 0.06172840 0.02298851 0.11842105 0.53571429 0.91111111
## [7] 0.45762712 0.62962963 1.54285714 1.60606061 2.46153846 1.66666667
## [13] 4.17647059 4.56250000 2.82608696 4.00000000 3.30000000 10.71428571
## [19] 8.00000000 4.56250000
```

```
require(lattice)
xyplot (Dose ~ odds_tumor, data = ex2116)
```



```
binResponse <- cbind (Tumor, Total-Tumor)
glm2_1 <- glm (binResponse ~ Dose, data=ex2116, family = binomial)
summary (glm2_1)
```

```
##
## Call:
## glm(formula = binResponse ~ Dose, family = binomial, data = ex2116)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -7.3577  -4.0473  -0.1515   2.9109   4.7729
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.86705    0.07673  -11.3   <2e-16 ***
## Dose         14.33377    0.93695   15.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 667.20  on 19  degrees of freedom
## Residual deviance: 277.05  on 18  degrees of freedom
## AIC: 368.44
##
## Number of Fisher Scoring iterations: 5
```

```
ex2116$cum_rel_tumor <- cumsum(Tumor)/sum(Tumor)
ex2116$cum_rel_tumor
```

```
## [1] 0.009473684 0.014736842 0.016842105 0.026315789 0.057894737 0.101052632
## [7] 0.129473684 0.165263158 0.222105263 0.277894737 0.345263158 0.403157895
## [13] 0.477894737 0.554736842 0.623157895 0.698947368 0.768421053 0.847368421
## [19] 0.923157895 1.000000000
```

Here, the p-value is less than 0.05. So, we can conclude that there are strong relationship between dose of Aflatoxin and odds of liver tumor. From the above output we can see that at dose 0.10, 50% of the fish will get liver tumors.

Exercise 22.15

```
ex2215 = case2201
attach(ex2215)
```

a

```
lm3_1 <- lm (sqrt(Matings) ~ Age)
summary (lm3_1)
```

```
##
## Call:
## lm(formula = sqrt(Matings) ~ Age)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.90532 -0.33654  0.07767  0.45871  1.09468
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.81220    0.56867  -1.428  0.161187
## Age          0.06320    0.01561   4.049  0.000236 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6493 on 39 degrees of freedom
## Multiple R-squared:  0.296, Adjusted R-squared:  0.2779
## F-statistic: 16.4 on 1 and 39 DF, p-value: 0.0002362
```

The fitted simple linear regression model for the square root of the matings is given by,

$$\hat{y}_i = -0.81220 + 0.06320(Age)$$

with $SE(\beta_0) = 0.56867$ and $SE(\beta_1) = 0.01561$

b

```
lm3_2 <- lm (log(Matings+1) ~ Age)
summary (lm3_2)
```

```
##
## Call:
## lm(formula = log(Matings + 1) ~ Age)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.49087 -0.33939  0.06607  0.35376  0.81171
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.69893    0.45861  -1.524 0.135567
## Age          0.05093    0.01259   4.046 0.000238 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5237 on 39 degrees of freedom
## Multiple R-squared:  0.2957, Adjusted R-squared:  0.2776
## F-statistic: 16.37 on 1 and 39 DF,  p-value: 0.0002385
```

The fitted simple linear regression model for the logarithmic transformation of the matings(after adding 1) is given by,

$$\log(\hat{y}_i) = -0.69893 + 0.05093(Age)$$

with $SE(\beta_0) = 0.45861$ and $SE(\beta_1) = 0.01259$

c

```
glm3_3 <- glm (Matings ~ Age, family = poisson)
summary (glm3_3)

##
## Call:
## glm(formula = Matings ~ Age, family = poisson)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.80798  -0.86137  -0.08629   0.60087   2.17777
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.58201    0.54462  -2.905  0.00368 **
## Age          0.06869    0.01375   4.997 5.81e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##      Null deviance: 75.372  on 40  degrees of freedom
## Residual deviance: 51.012  on 39  degrees of freedom
## AIC: 156.46
##
## Number of Fisher Scoring iterations: 5
```

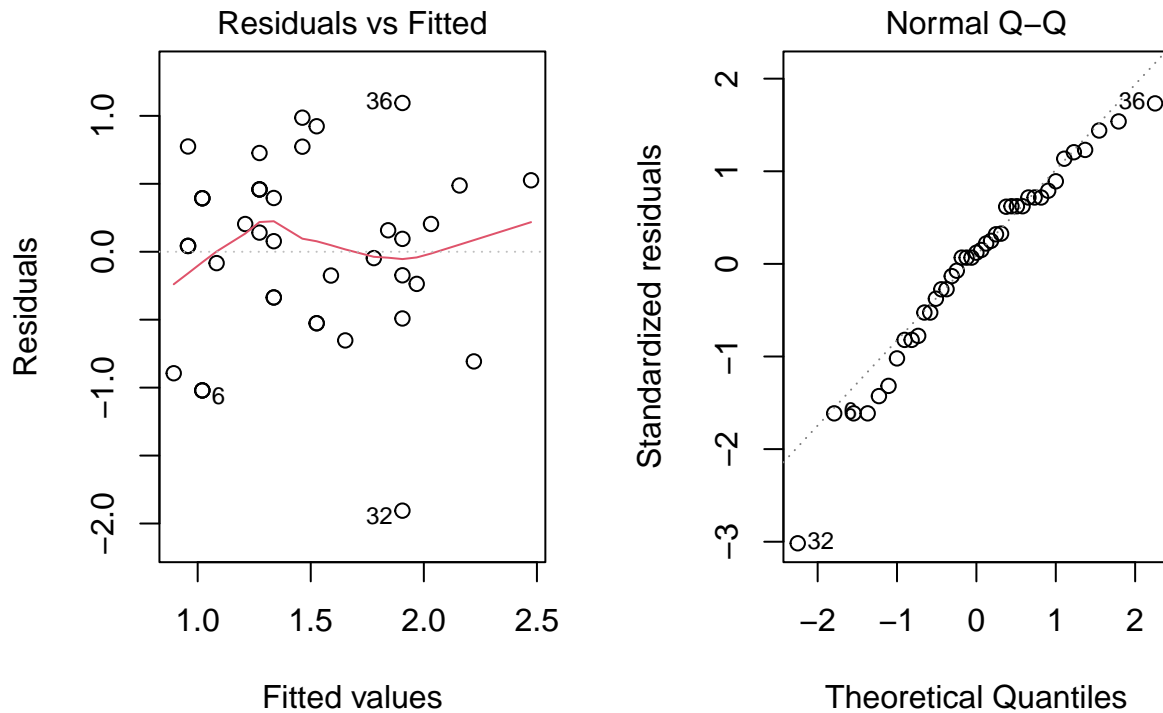
With a count response matings and a explanatory variable age, the Poisson log-linear regression model is given by,

$$\log(\hat{\mu}) = -1.58201 + 0.06869(Age)$$

with $SE(\beta_0) = 0.54462$ and $SE(\beta_1) = 0.01375$

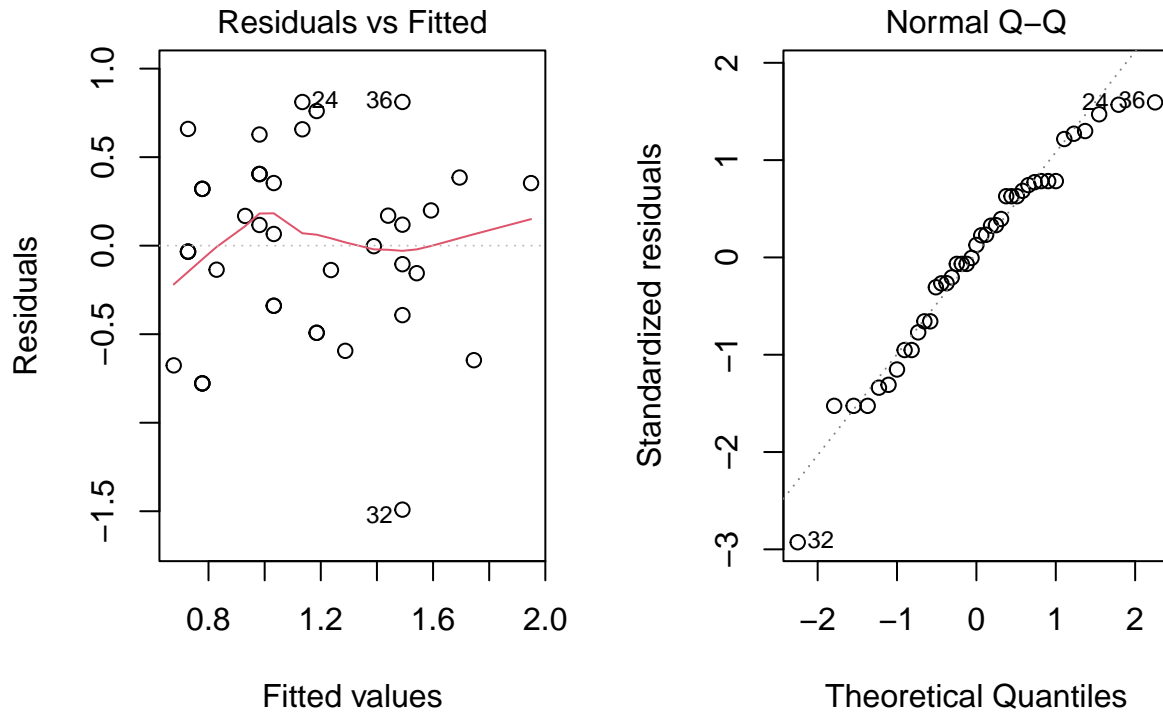
d

```
par(mfrow = c(1, 2))
plot(lm3_1, 1)
plot(lm3_1, 2)
```



The residual plot will show no fitted pattern. That is, the red line should be approximately horizontal at zero. But in that case of model fitted in (a), the red line is not horizontal at 0. In case of the normal probability plot of residuals should approximately follow a straight line which indicate the model is adequate but in our case for the model fitted in (a) doesn't show the straight line rather showing some outliers at case no 32, and 36. So from both the plot we can assume that the fitted model in (a) exhibit obvious inadequacy.

```
par(mfrow = c(1, 2))
plot(lm3_2, 1)
plot(lm3_2, 2)
```

The fitted plot seems perfect to me.

Exercise 22.22

```
attach (ex2222)
glm5_1 <- glm (Count ~ Book + Word, family = poisson)
summary (glm5_1)
```

```
##
## Call:
## glm(formula = Count ~ Book + Word, family = poisson)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -3.3545  -1.0360  -0.2218   1.0895   3.2864
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      5.23397    0.05816  89.993  <2e-16 ***
## BookSanditonI     -0.77851    0.08499  -9.160  <2e-16 ***
## BookSanditonII    -0.80866    0.08588  -9.417  <2e-16 ***
## BookSense&Sensibility -0.15985    0.07028  -2.274   0.0229 *
## Wordan           -1.73718    0.11368 -15.281  <2e-16 ***
## Wordthat         -0.69508    0.07622  -9.119  <2e-16 ***
## Wordthis         -1.63292    0.10879 -15.010  <2e-16 ***
```

```
## Wordwith          -0.92992    0.08268 -11.247    <2e-16 ***
## Wordwithout       -2.51037    0.16045 -15.646    <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##      Null deviance: 878.858  on 23  degrees of freedom
## Residual deviance:  44.324  on 15  degrees of freedom
## AIC: 190.91
##
## Number of Fisher Scoring iterations: 4
```

For this model, the p-value for the coefficients are less than the critical value (0.05) at 5% level of significance which indicates that Jane Austen was consistent in the three books in her relative uses of these words.

```
ind_sandII <- ifelse (Book == "SanditonII", 1, 0)
glm5_2 <- glm (Count ~ ind_sandII + Word, family = poisson)
summary (glm5_2)
```

```
##
## Call:
## glm(formula = Count ~ ind_sandII + Word, family = poisson)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -4.5646  -1.5535   0.0266   1.7490   3.6245
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.97319    0.04575 108.702 < 2e-16 ***
## ind_sandII   -0.54789    0.07801  -7.023 2.16e-12 ***
## Wordan       -1.73718    0.11368 -15.281 < 2e-16 ***
## Wordthat     -0.69508    0.07622  -9.119 < 2e-16 ***
## Wordthis     -1.63292    0.10879 -15.010 < 2e-16 ***
## Wordwith     -0.92992    0.08268 -11.247 < 2e-16 ***
## Wordwithout  -2.51037    0.16045 -15.646 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##      Null deviance: 878.86  on 23  degrees of freedom
## Residual deviance: 140.34  on 17  degrees of freedom
## AIC: 282.92
##
## Number of Fisher Scoring iterations: 4
```

The emulator do a great job in terms of matching the relative rates of occurrence of these six words since the two-sided p-value is smaller than the critical value (0.05) at 5% level of significance.

```
ind_sandI <- ifelse (Book == "SanditonI", 1, 0)
glm5_3 <- glm (Count ~ ind_sandI + ind_sandII + Word, family = poisson)
summary (glm5_3)
```

```
##
```

```

## Call:
## glm(formula = Count ~ ind_sandI + ind_sandII + Word, family = poisson)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -3.3545  -1.0734  -0.3329   0.9411   3.2864
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   5.15723    0.04834 106.685  <2e-16 ***
## ind_sandI     -0.70177    0.07860  -8.929  <2e-16 ***
## ind_sandII    -0.73193    0.07956  -9.200  <2e-16 ***
## Wordan        -1.73718    0.11368 -15.281  <2e-16 ***
## Wordthat      -0.69508    0.07622  -9.119  <2e-16 ***
## Wordthis      -1.63292    0.10879 -15.010  <2e-16 ***
## Wordwith      -0.92992    0.08268 -11.247  <2e-16 ***
## Wordwithout   -2.51037    0.16045 -15.646  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##      Null deviance: 878.858  on 23  degrees of freedom
## Residual deviance:  49.514  on 16  degrees of freedom
## AIC: 194.1
##
## Number of Fisher Scoring iterations: 4

```

From the above output we can conclude that, the emulator match the relative rates that Austen used the words in the first part of Sanditon since the two-sided p-value is smaller than the critical value (0.05) at 5% level of significance.