Awards

Yang

February 11, 2019

```r
awards<-read.csv("awards.csv")
attach(awards)
```

Comments on significance of predictors

Model 1: Math as a continuous predictor

Fit

```r
m1=glm(numawards~1+math,family = poisson(link = log),x=TRUE)
summary(m1)
```

```
##
## Call:
## glm(formula = numawards ~ 1 + math, family = poisson(link = log),
##     x = TRUE)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1853  -0.9070  -0.6001   0.3246   2.9529
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.333532   0.591261  -9.021   <2e-16 ***
## math         0.086166   0.009679   8.902   <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: 287.67  on 199  degrees of freedom
## Residual deviance: 204.02  on 198  degrees of freedom
## AIC: 384.08
##
## Number of Fisher Scoring iterations: 6
```

Comment on the significance of math as a predictor

Math is a significant predictor for number of awards since the p-value is small(<0.05).

Model 2: Prog as a factor predictor

Fit

```
m2=glm(numawards~1+as.factor(prog),family=poisson(link=log),x=TRUE)
summary(m2)
```

```
##
## Call:
## glm(formula = numawards ~ 1 + as.factor(prog), family = poisson(link = log),
##     x = TRUE)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4142  -0.6928  -0.6325   0.0000   3.3913
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.6094     0.3333  -4.828 1.38e-06 ***
## as.factor(prog)2   1.6094     0.3473   4.634 3.59e-06 ***
## as.factor(prog)3   0.1823     0.4410   0.413    0.679
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 287.67  on 199  degrees of freedom
## Residual deviance: 234.46  on 197  degrees of freedom
## AIC: 416.51
##
## Number of Fisher Scoring iterations: 6
```

Comments on significance of params

The model has 3 parameters. And β_0 , β_1 are statistically important since they have small p-value(1.38e-06 and 3.59e-06 respectively) compared with 0.05 while β_2 is not statistically significant in our model since it has a large p-value 0.679.

Wald test

```
I = t(m2$x)%*%diag(m2$weights)%*%m2$x
I_inv = solve(I)
sd <- sqrt(diag(I_inv))
p_value <- pchisq((m2$coefficients/sd)^2,df=2,lower.tail=FALSE)
p_value
```

```
##      (Intercept) as.factor(prog)2 as.factor(prog)3
##     8.664234e-06     2.174618e-05     9.180740e-01
```

Meanwhile, according to the wald test, factor2("Academic") is statistically significant with p-value 2.174618e-05 while factor3("Vocational") is not statistically significant due to its p-value 9.180740e-01.

Likelihood ratio test

```
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 3.4.4
## Loading required package: zoo
## Warning: package 'zoo' was built under R version 3.4.4
##
## Attaching package: 'zoo'
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
m_null=glm(numawards~1,family=poisson(link=log),x=TRUE)
test=lrtest(m2,m_null)
test
```

```
## Likelihood ratio test
##
## Model 1: numawards ~ 1 + as.factor(prog)
## Model 2: numawards ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   3 -205.26
## 2   1 -231.86 -2 53.212  2.787e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
test$`Pr(>Chisq)`
```

```
## [1]           NA 2.786791e-12
```

According to the likelihood ratio test resulting in p-value=2.786791e-12<0.05, we reject the null model and can conclude that prog is a significant predictor.

Model 3: numawards~1+math+as.factor(prog)

Fit

```
m3=glm(numawards~1+math+as.factor(prog), family=poisson(link='log'),x=TRUE)
summary(m3)
```

```
##
## Call:
## glm(formula = numawards ~ 1 + math + as.factor(prog), family = poisson(link = "log"),
##     x = TRUE)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2043  -0.8436  -0.5106   0.2558   2.6796
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -5.24712    0.65845  -7.969 1.60e-15 ***
```

```
## math               0.07015    0.01060   6.619 3.63e-11 ***
## as.factor(prog)2  1.08386    0.35825   3.025  0.00248 **
## as.factor(prog)3  0.36981    0.44107   0.838  0.40179
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 287.67  on 199  degrees of freedom
## Residual deviance: 189.45  on 196  degrees of freedom
## AIC: 373.5
##
## Number of Fisher Scoring iterations: 6
```

model 3:numawards~1+math+as.factor(prog), has 4 parameters.

Interpretation

Expected number of awards should increase by $e^{0.07015}$ if **math** is increased by 1.
The difference in expected number of awards between student enrolled in **academic** program and in general program is $e^{1.08386}$.
The difference in expected number of awards between student enrolled in **vocational** program and in general program is $e^{0.36981}$.

Model 4: numawards~1 + math * as.factor(prog)

Fit

```r
m4=glm(numawards~1+math*as.factor(prog), family=poisson(link='log'),x=TRUE)
summary(m4)
```

```
##
## Call:
## glm(formula = numawards ~ 1 + math * as.factor(prog), family = poisson(link = "log"),
##     x = TRUE)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2295  -0.7958  -0.5298   0.2528   2.6826
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -3.86179    2.49317  -1.549    0.121
## math                   0.04400    0.04721   0.932    0.351
## as.factor(prog)2      -0.44107    2.60299  -0.169    0.865
## as.factor(prog)3      -0.84473    2.86990  -0.294    0.768
## math:as.factor(prog)2  0.02841    0.04870   0.583    0.560
## math:as.factor(prog)3  0.02290    0.05421   0.422    0.673
##
## (Dispersion parameter for poisson family taken to be 1)
```

```
##
##     Null deviance: 287.67  on 199  degrees of freedom
## Residual deviance: 189.10  on 194  degrees of freedom
## AIC: 377.16
##
## Number of Fisher Scoring iterations: 6
```

model 4: numawards ~ 1 + math * as.factor(prog), has 6 parameters.

Interpretation

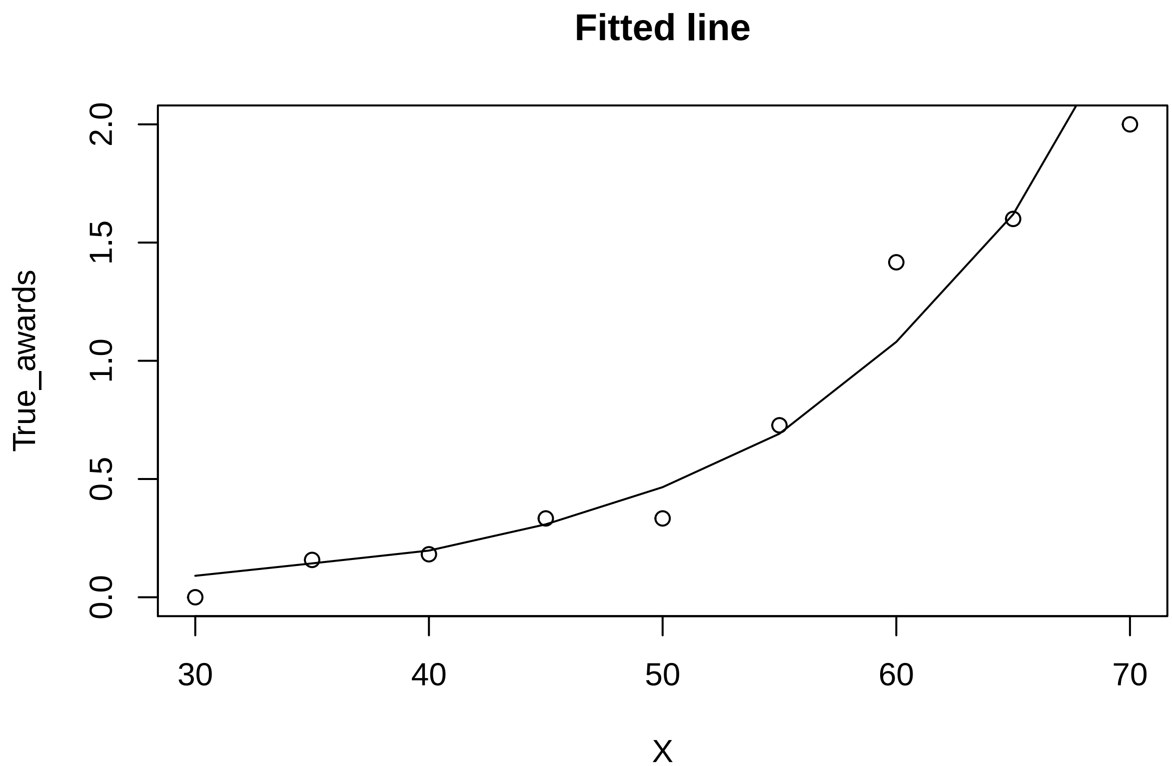
Expected number of awards should increase by $e^{0.044}$ if **math** is increased by 1.
The difference in expected number of awards between student enrolled in **academic** program and in general program is $e^{-0.44107}$.
The difference in expected number of awards between student enrolled in **vocational** program and in general program is $e^{-0.84473}$.
The difference in expected number of awards between student enrolled in **academic** program and in general program is expected to increase by $e^{0.02841}$ if **math** increase by 1.
The difference in expected number of awards between student enrolled in vocational program and in general program is expected to increase by $e^{0.02841}$ if **math** increase by 1.

Using Plot to Analyze

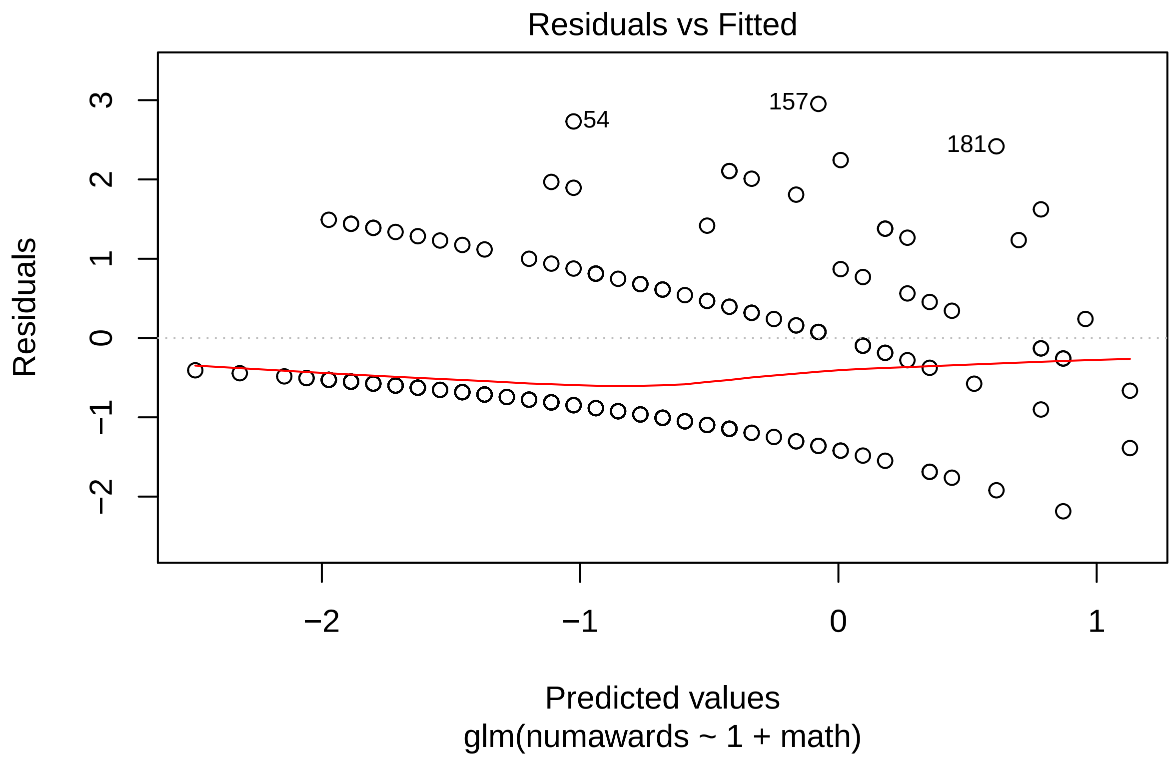
model1

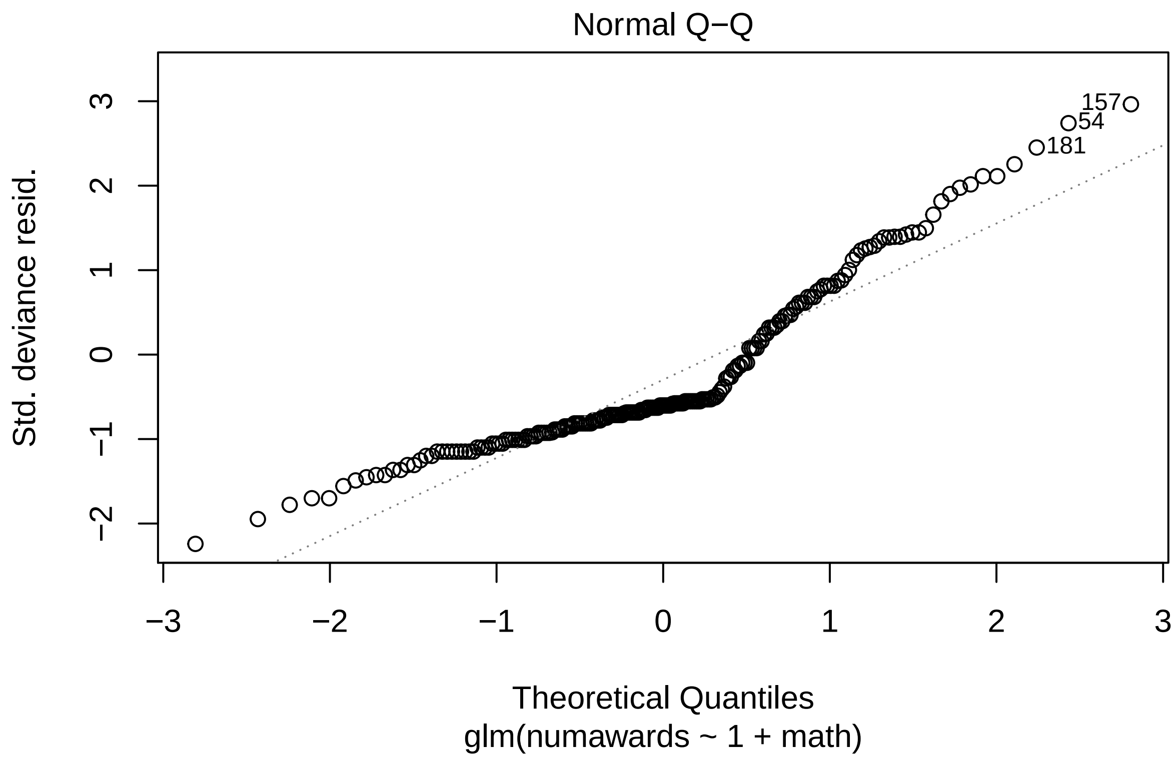
```r
awards$m1=m1$fitted.values
awards$m2=m2$fitted.values
awards$m3=m3$fitted.values
awards$m4=m4$fitted.values
X=seq(30,70,5)
df= split(awards, cut(awards$math, breaks = seq(30,75,5)))
Mean_award <- function(x){
  mean(x$numawards)
}
Mean_prediction_m1 <- function(x){
  mean(x$m1)
}
True_awards = sapply(df, Mean_award)
Estimated_awards_1=sapply(df,Mean_prediction_m1)

plot(X,True_awards,main="Fitted line")
lines(X,Estimated_awards_1)
```

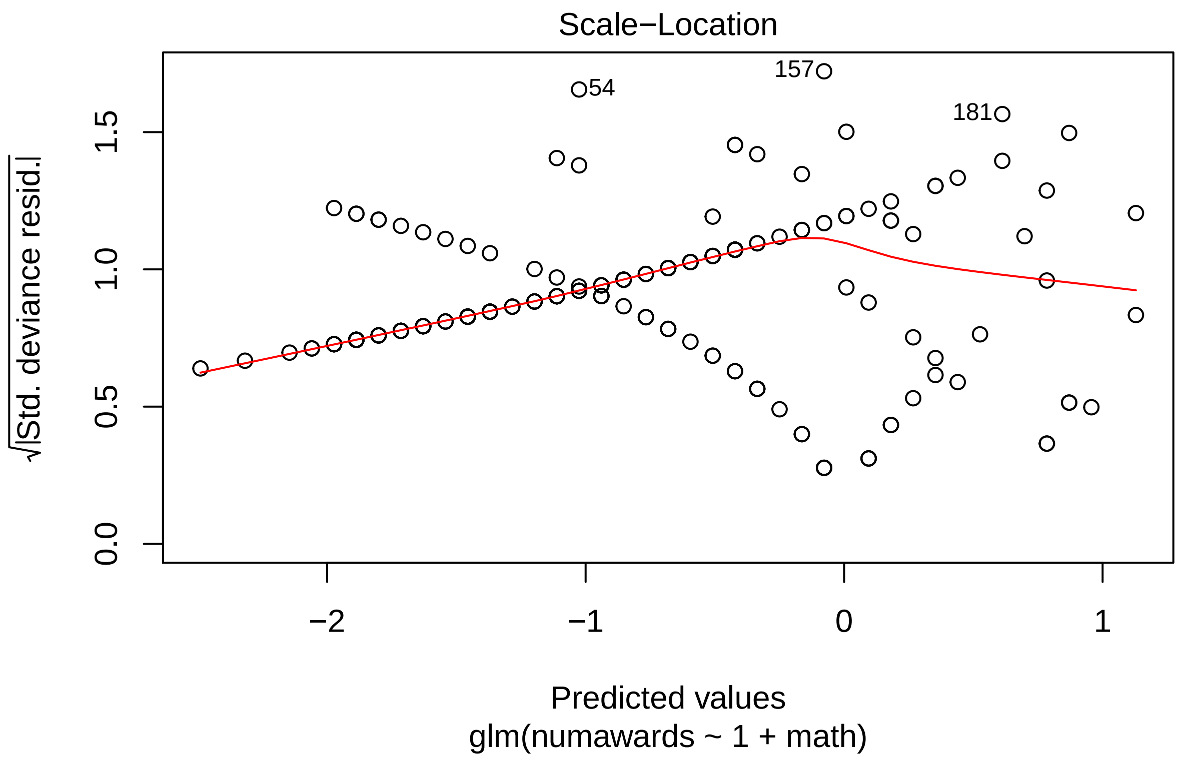


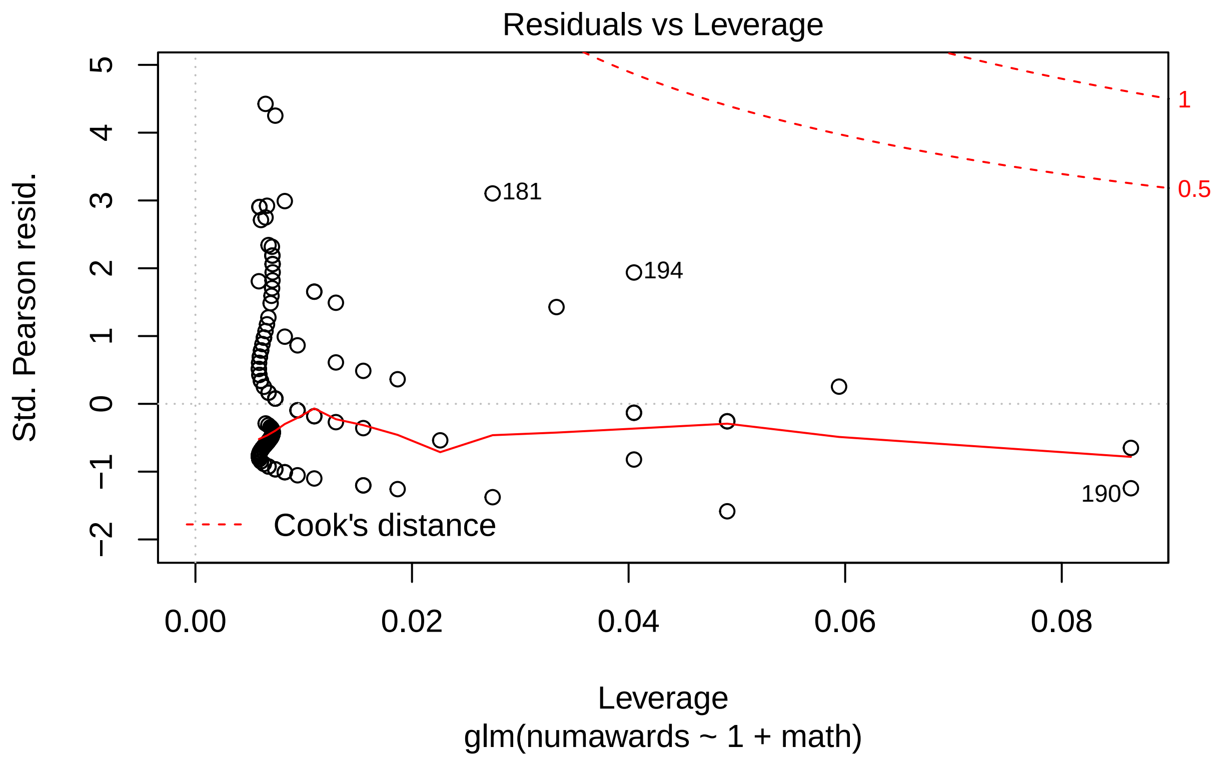

```r
plot(m1)
```


Residuals vs Fitted
Residuals
Predicted values
glm(numawards ~ 1 + math)
3
2
1
0
−1
−2
−2
−1
0
1
54
157
181



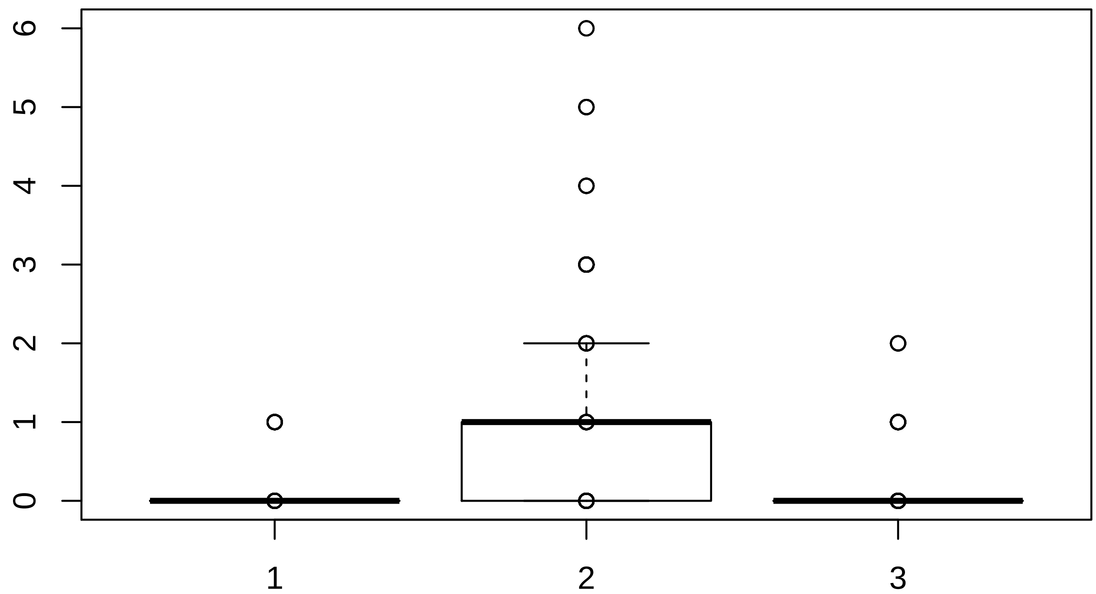
Normal Q−Q
Std. deviance resid.
3
2
1
0
−1
−2
157
54
181
−3
−2
−1
0
1
2
3
Theoretical Quantiles
glm(numawards ~ 1 + math)


Scale−Location
157
54
181
1.5
1.0
0.5
0.0
√|Std. deviance resid.|
−2
−1
0
1
Predicted values
glm(numawards ~ 1 + math)

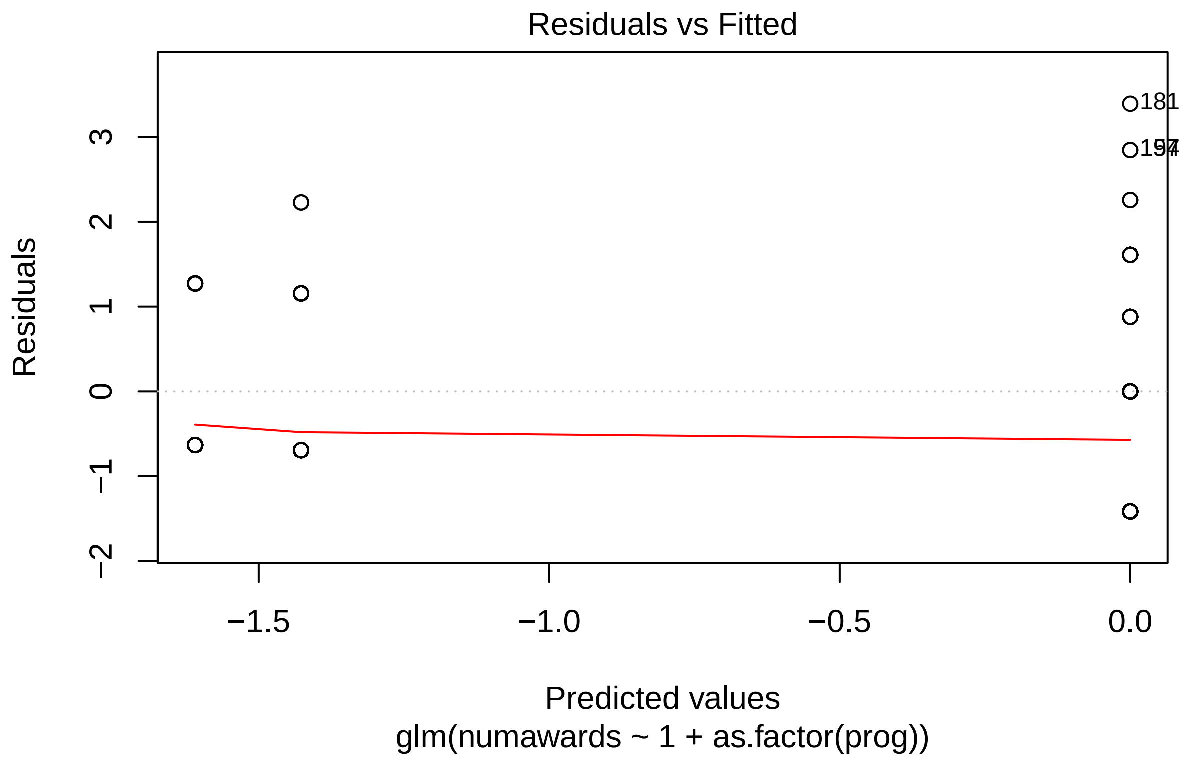


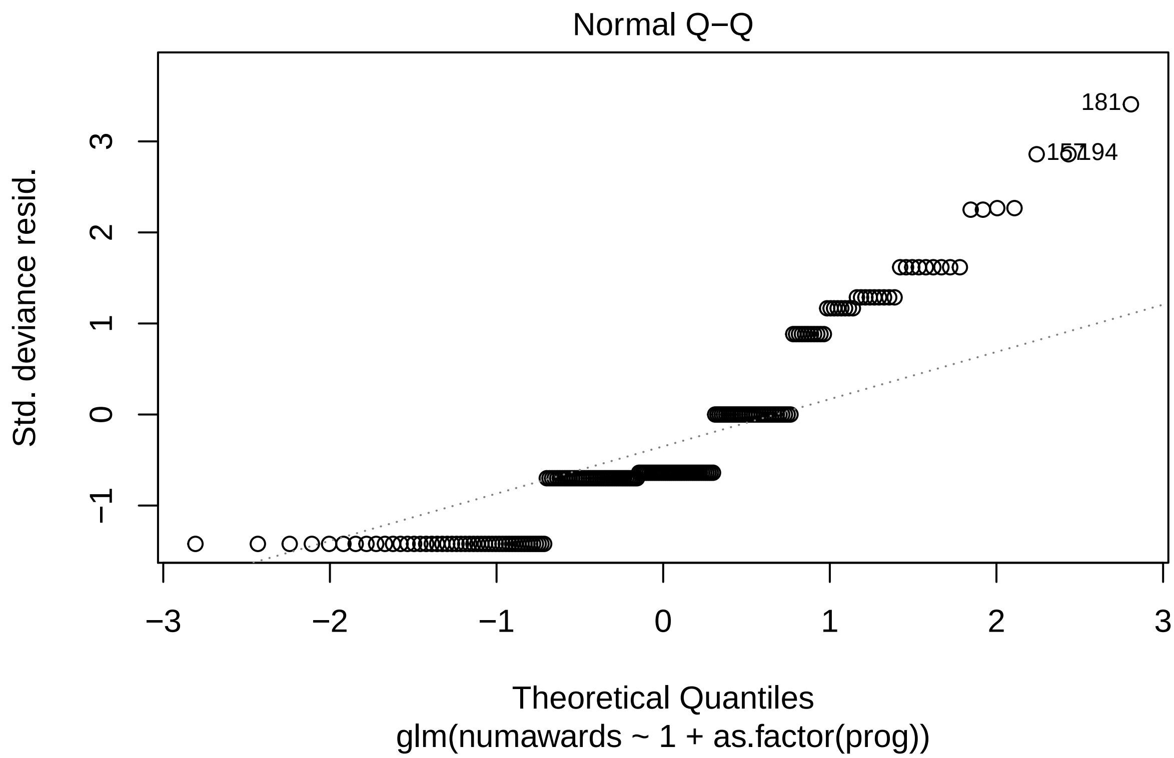
model2

```
Mean_prediction_m2 <- function(x){
  mean(x$m2)
}
Estimated_awards_2=sapply(df,Mean_prediction_m2)
plot(as.factor(awards$prog),awards$numawards)
academic=subset(awards,prog==2)
vocational=subset(awards,prog==3)
general=subset(awards,prog==1)
points(awards$prog,awards$numawards)
```

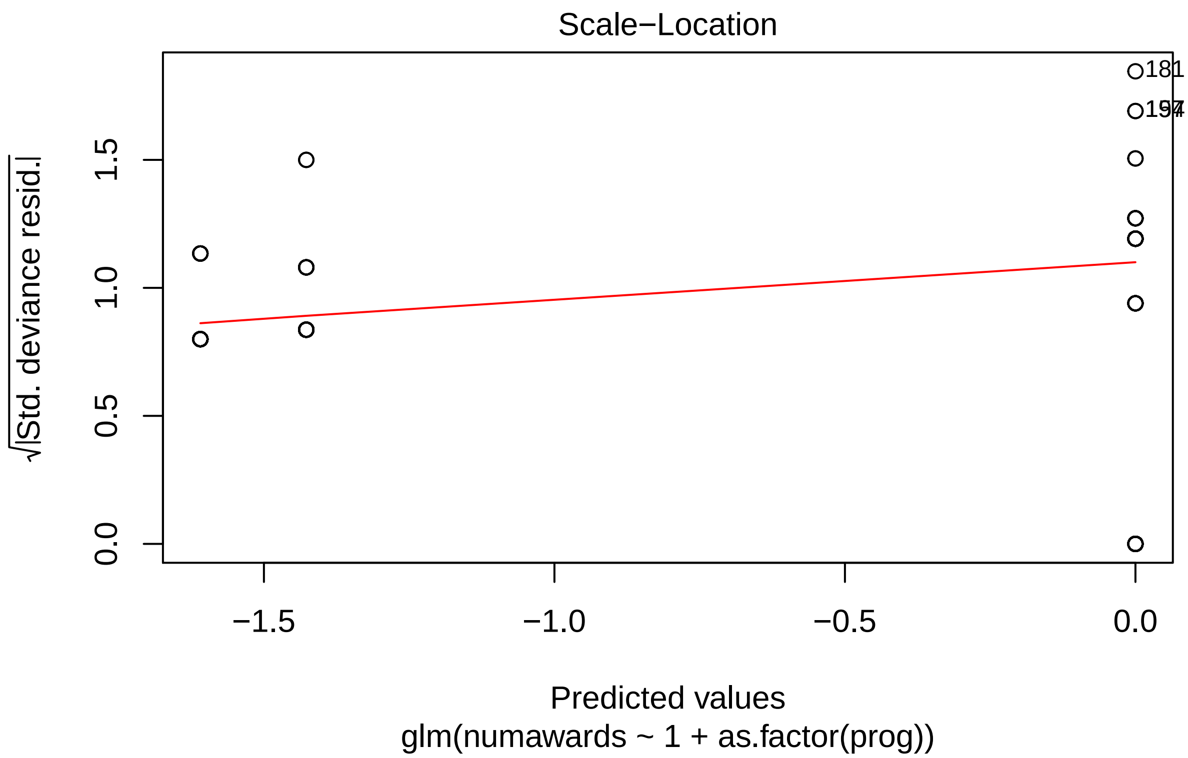



```r
plot(m2)
```

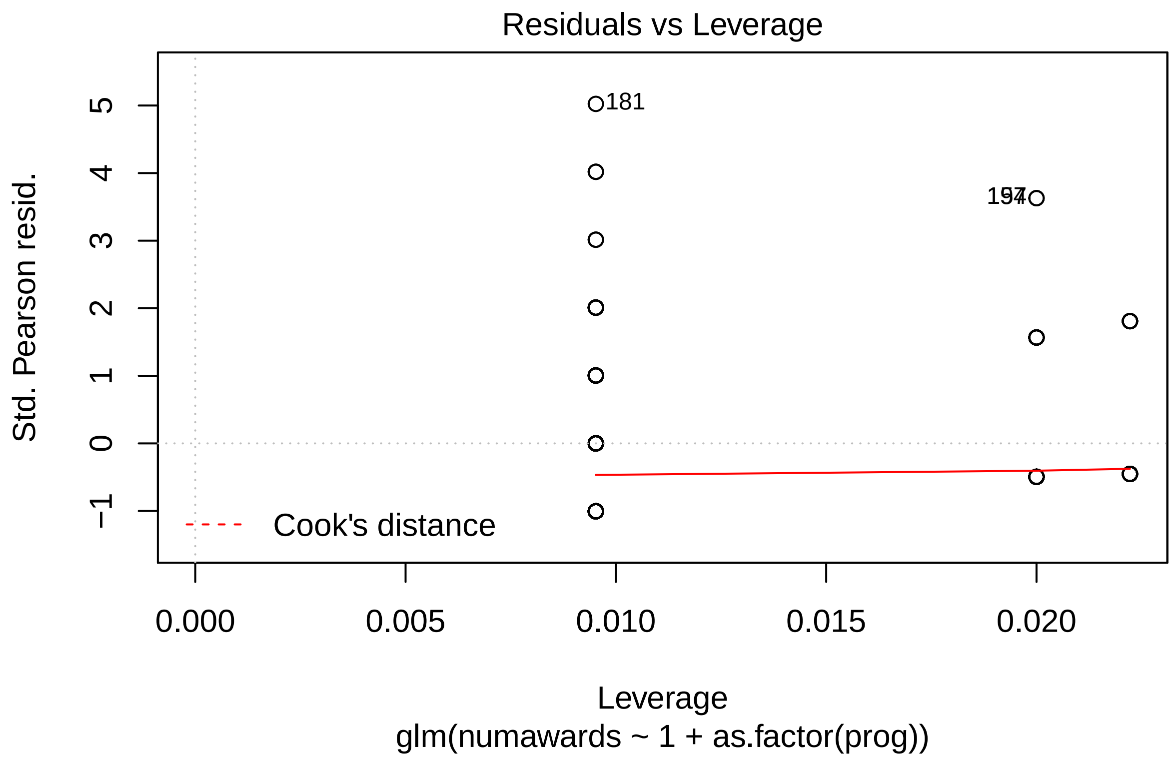

Residuals vs Fitted
Residuals
3
2
1
0
−1
−2
181
197
54
−1.5
−1.0
−0.5
0.0
Predicted values
glm(numawards ~ 1 + as.factor(prog))



Normal Q−Q
Std. deviance resid.
Theoretical Quantiles
glm(numawards ~ 1 + as.factor(prog))
181
157
194
−3
−2
−1
0
1
2
3



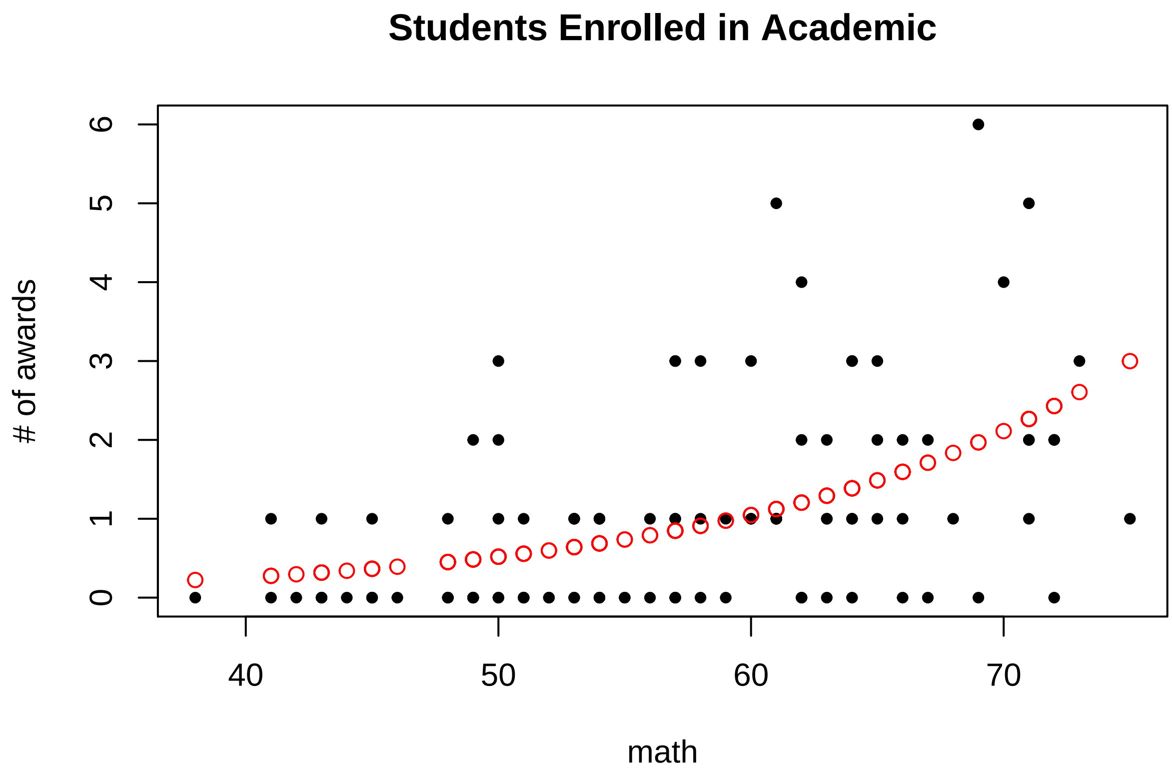
Scale−Location
√|Std. deviance resid.|
0.0
0.5
1.0
1.5
181
194
97
−1.5
−1.0
−0.5
0.0
Predicted values
glm(numawards ~ 1 + as.factor(prog))



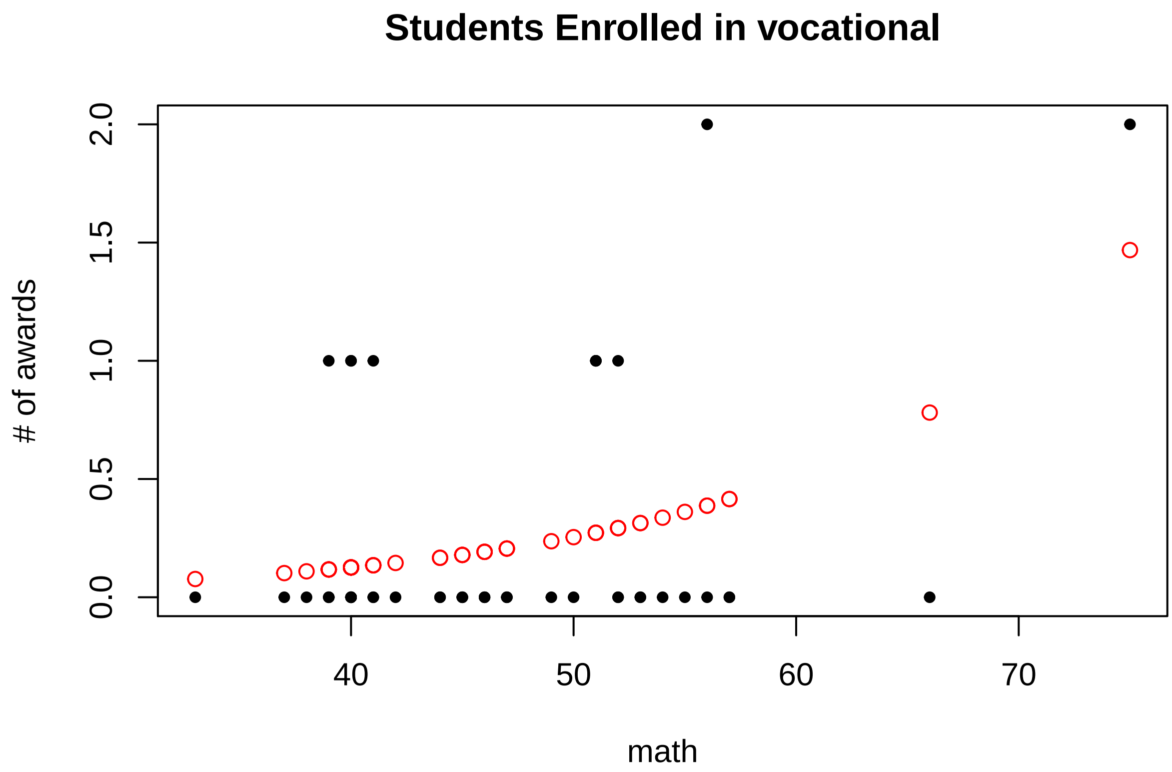
model3

The black dots are true datapoints and the red dots are # of awards estimated by model 3.

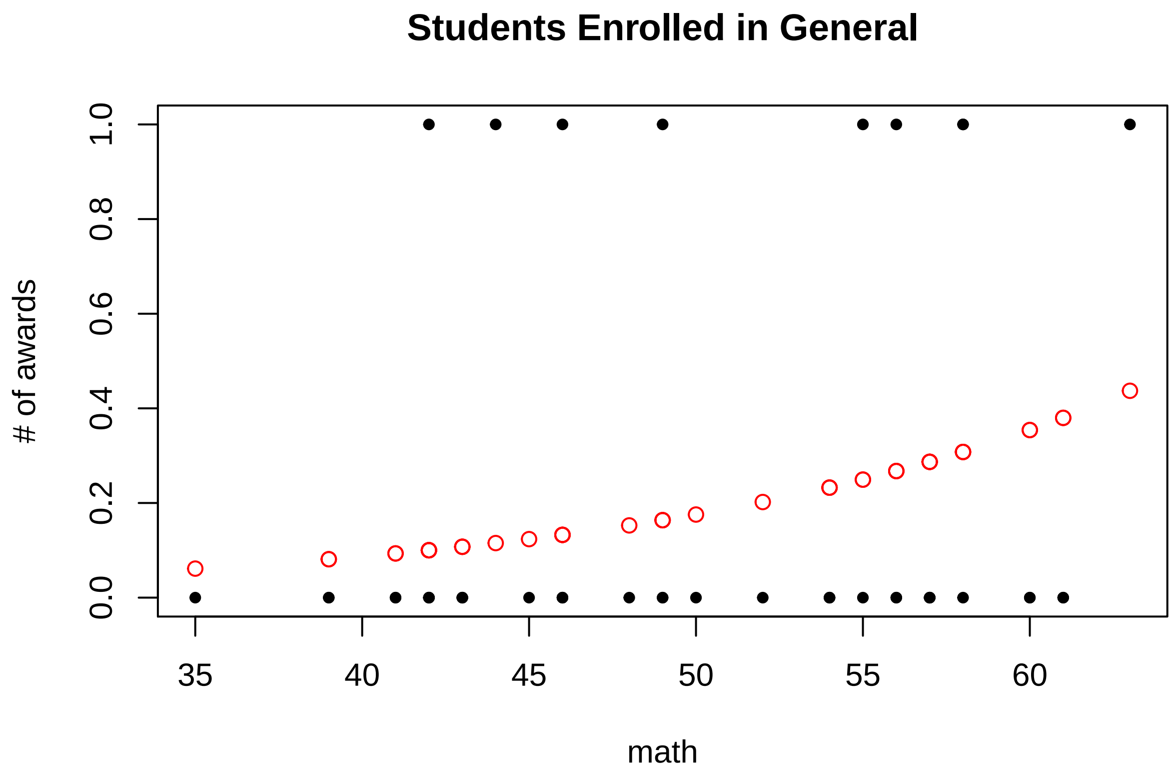
```
plot(academic$math,academic$numawards,col="black",pch=20,xlab="math",ylab = "# of awards",main = "Studen
points(academic$math,academic$m3,col="red")
legend(x=3.5,y=14, legend=c("True", "Predicted"),fill=c("black", "red"))
```



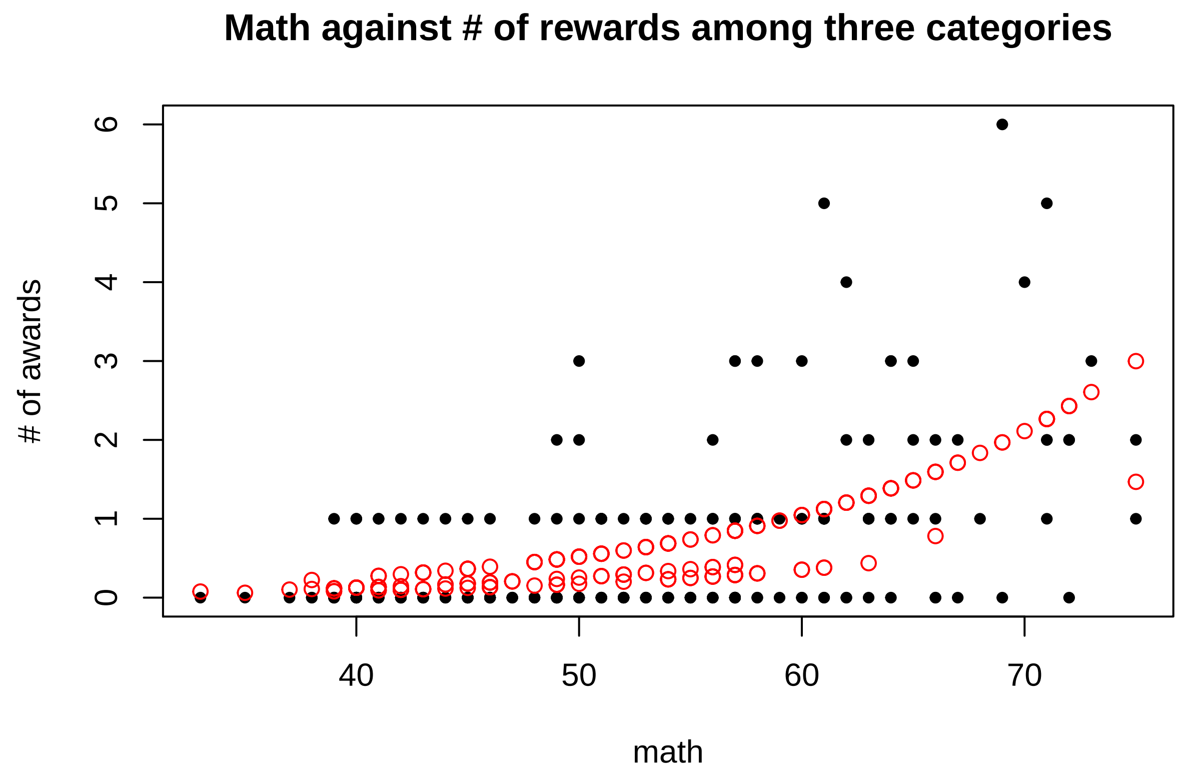

```
plot(vocational$math,vocational$numawards,col="black",pch=20,xlab="math",ylab = "# of awards",main = "St
points(vocational$math,vocational$m3,col="red")
legend(x=3.5,y=14, legend=c("True", "Predicted"),fill=c("black", "red"))
```



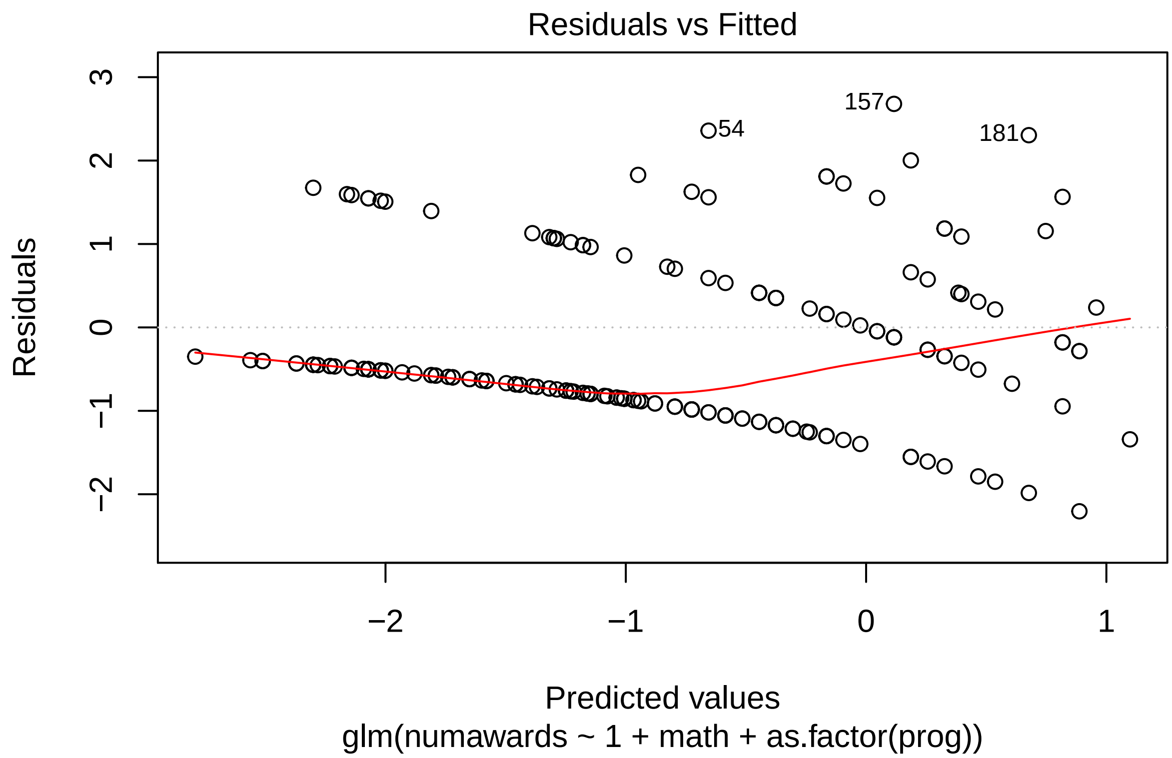

```
plot(general$math,general$numawards,col="black",pch=20,xlab="math",ylab = "# of awards",main = "Students
points(general$math,general$m3,col="red")
legend(x=3.5,y=14, legend=c("True", "Predicted"),fill=c("black", "red"))
```

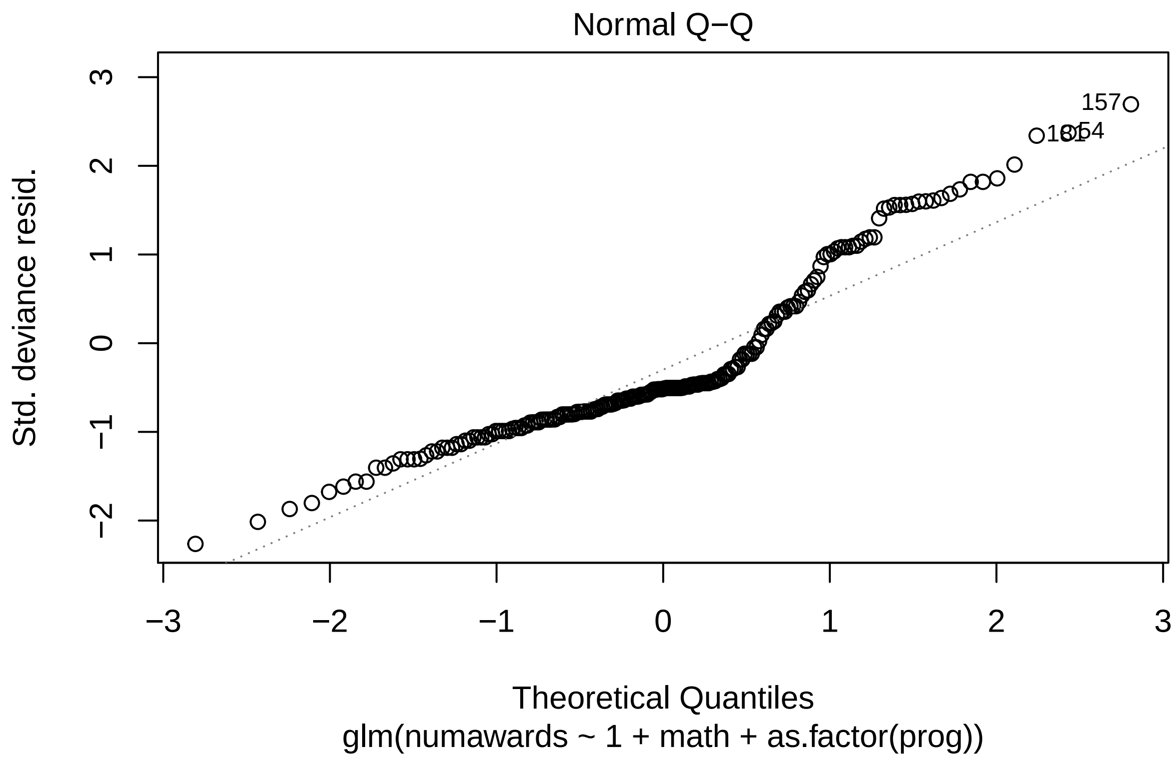


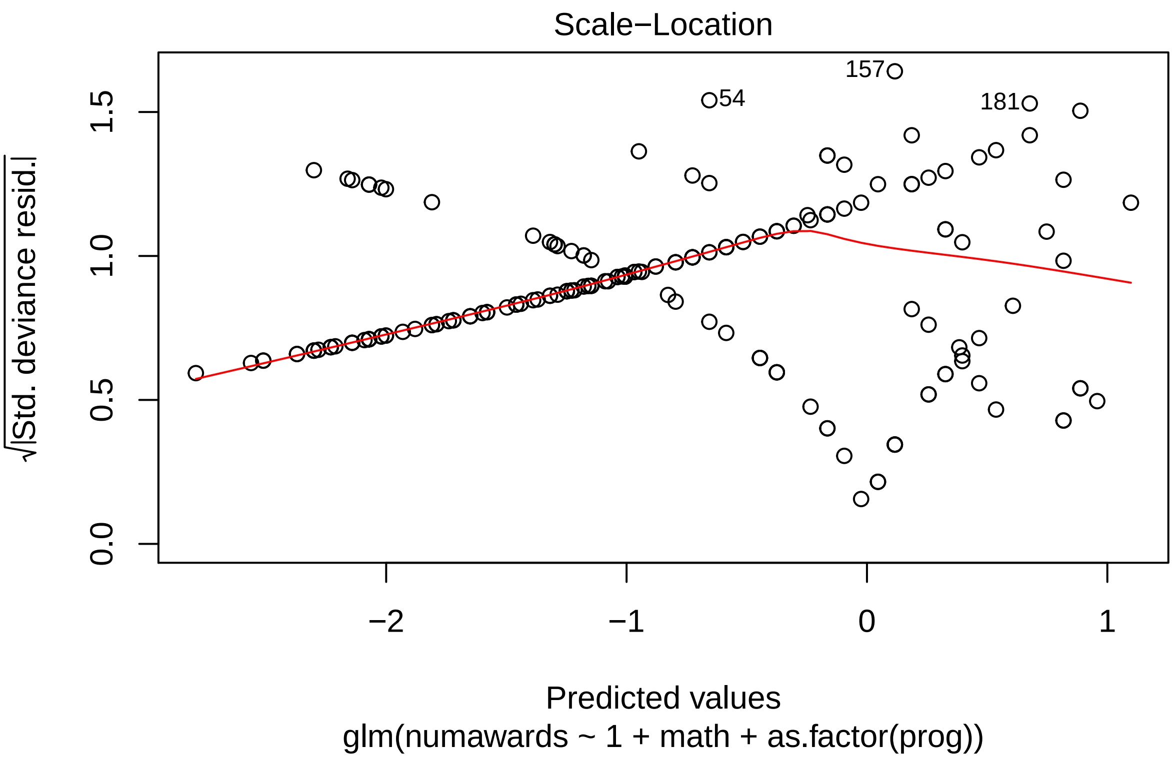

```
plot(awards$math,awards$numawards,col="black",pch=20,xlab="math",ylab = "# of awards",main = "Math agai
points(awards$math,awards$m3,col="red")
legend(x=3.5,y=14, legend=c("True", "Predicted"),fill=c("black", "red"))
```


Math against # of rewards among three categories
# of awards
0
1
2
3
4
5
6
40
50
60
70
math

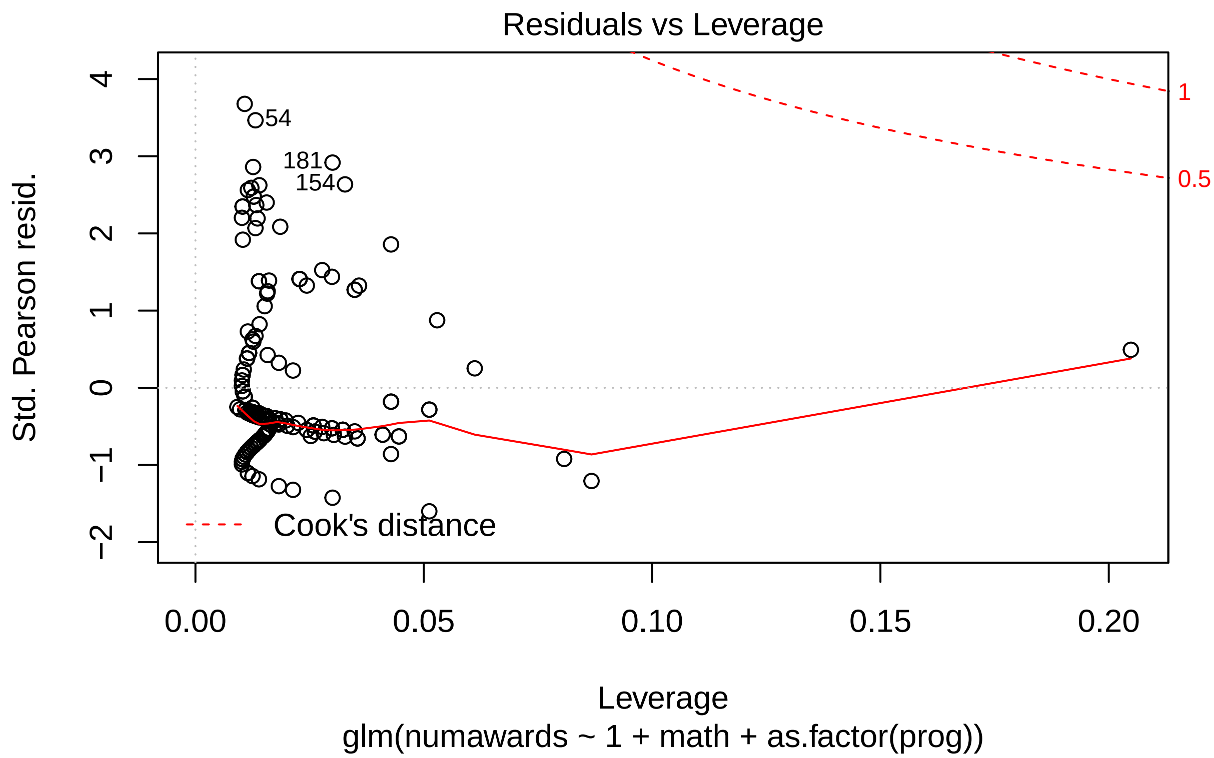
```r
plot(m3)
```


Residuals vs Fitted
Residuals
Predicted values
glm(numawards ~ 1 + math + as.factor(prog))
3
2
1
0
−1
−2
−2
−1
0
1
157
54
181


Normal Q−Q
Std. deviance resid.
Theoretical Quantiles
glm(numawards ~ 1 + math + as.factor(prog))
157
154
181
−3
−2
−1
0
1
2
3



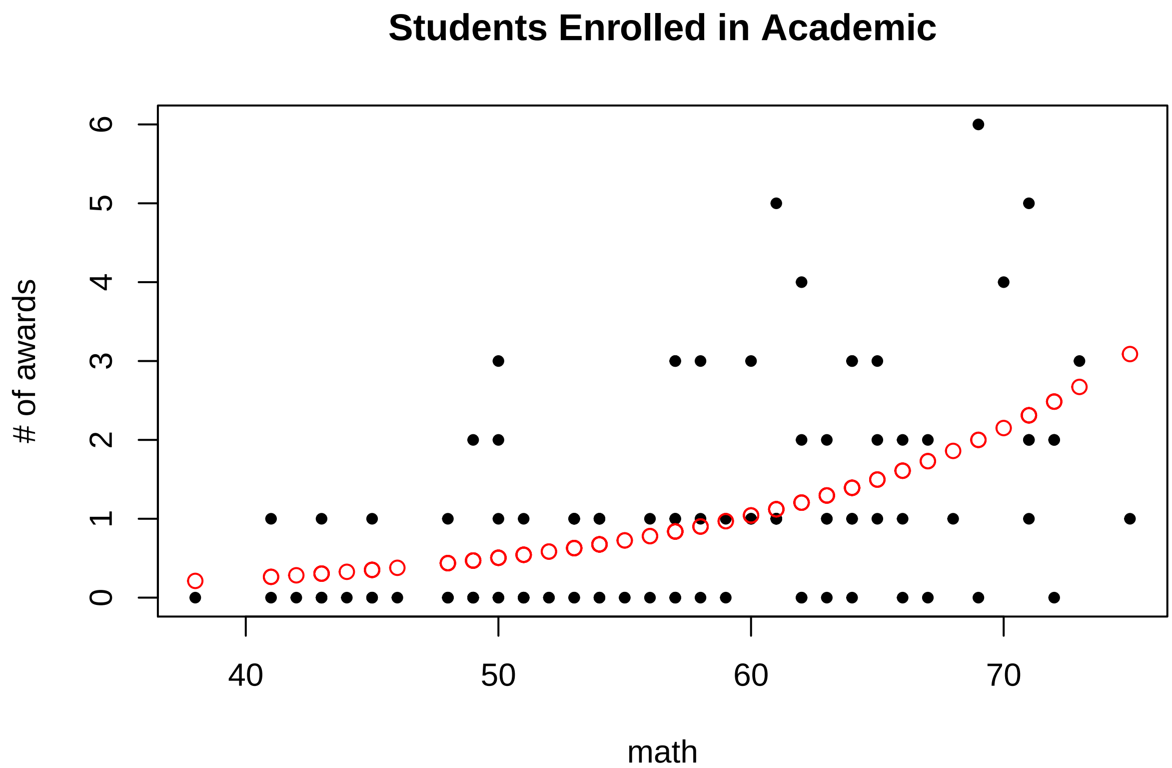
Scale−Location
157
54
181
1.5
1.0
0.5
0.0
√|Std. deviance resid.|
−2
−1
0
1
Predicted values
glm(numawards ~ 1 + math + as.factor(prog))



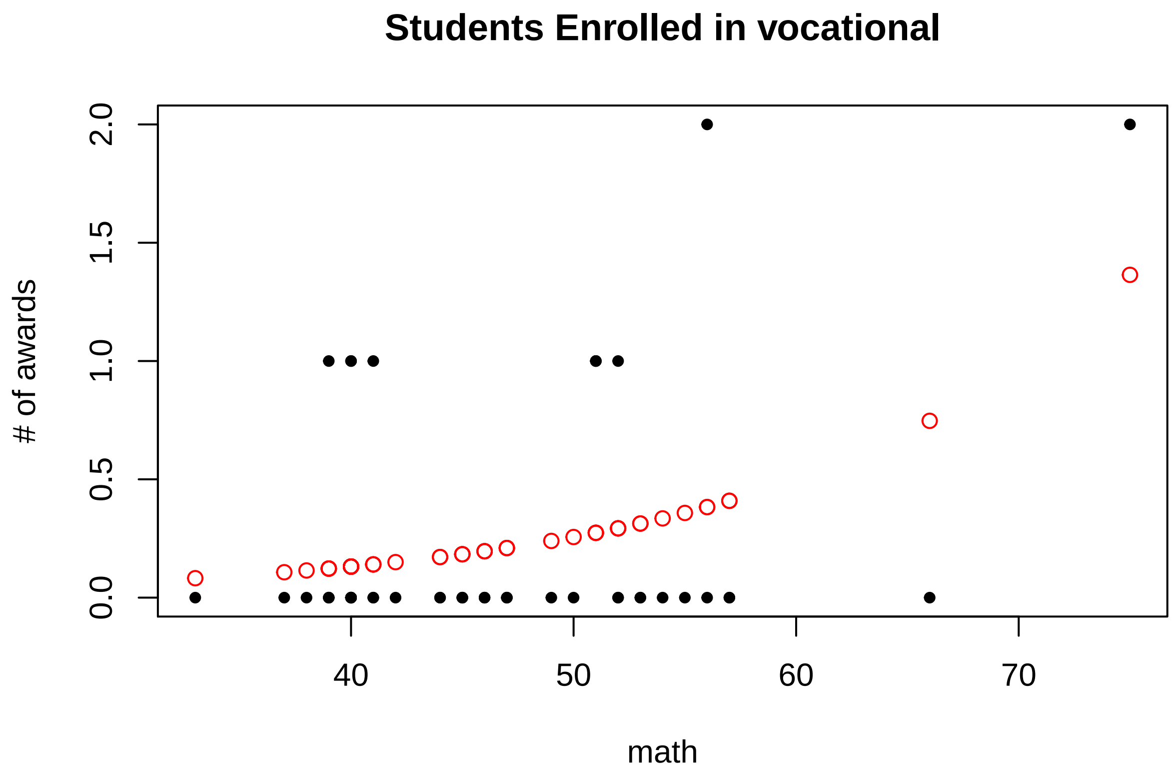
model4

Note: The black dots are true datapoints and the red dots are # of awards estimated by model 4.

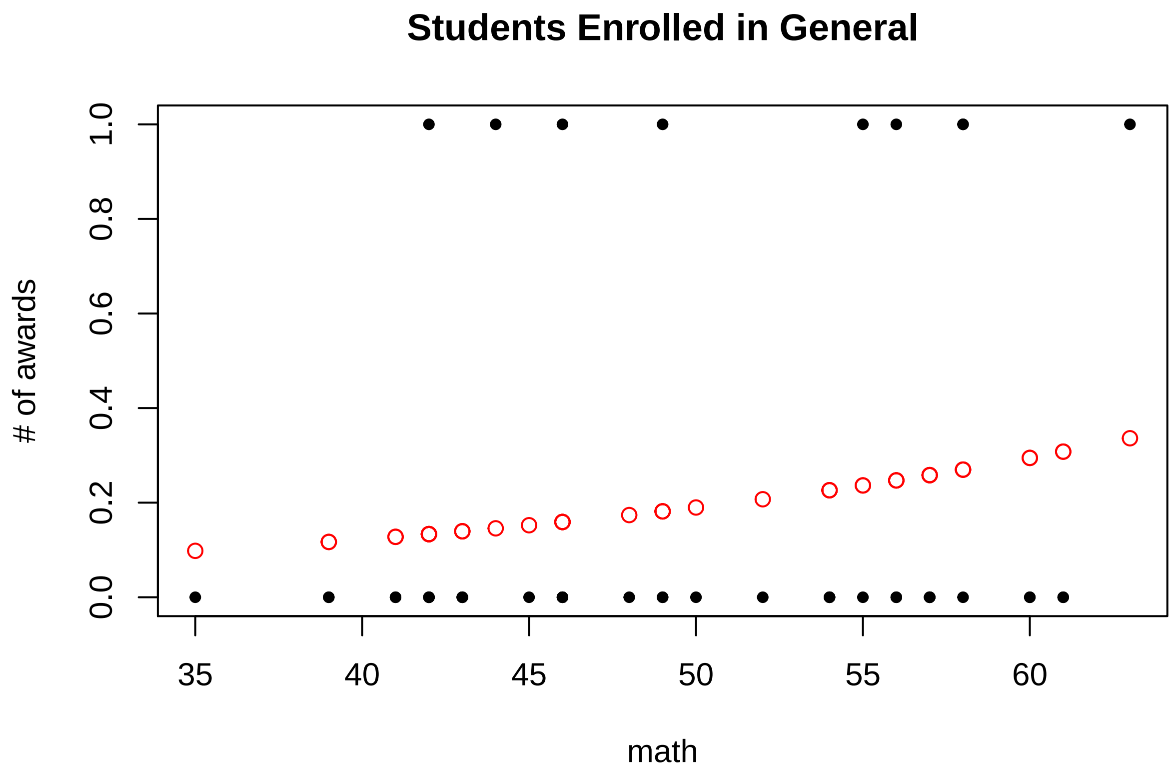
```
plot(academic$math,academic$numawards,col="black",pch=20,xlab="math",ylab = "# of awards",main = "Studen
points(academic$math,academic$m4,col="red")
legend(x=3.5,y=14, legend=c("True", "Predicted"),fill=c("black", "red"))
```



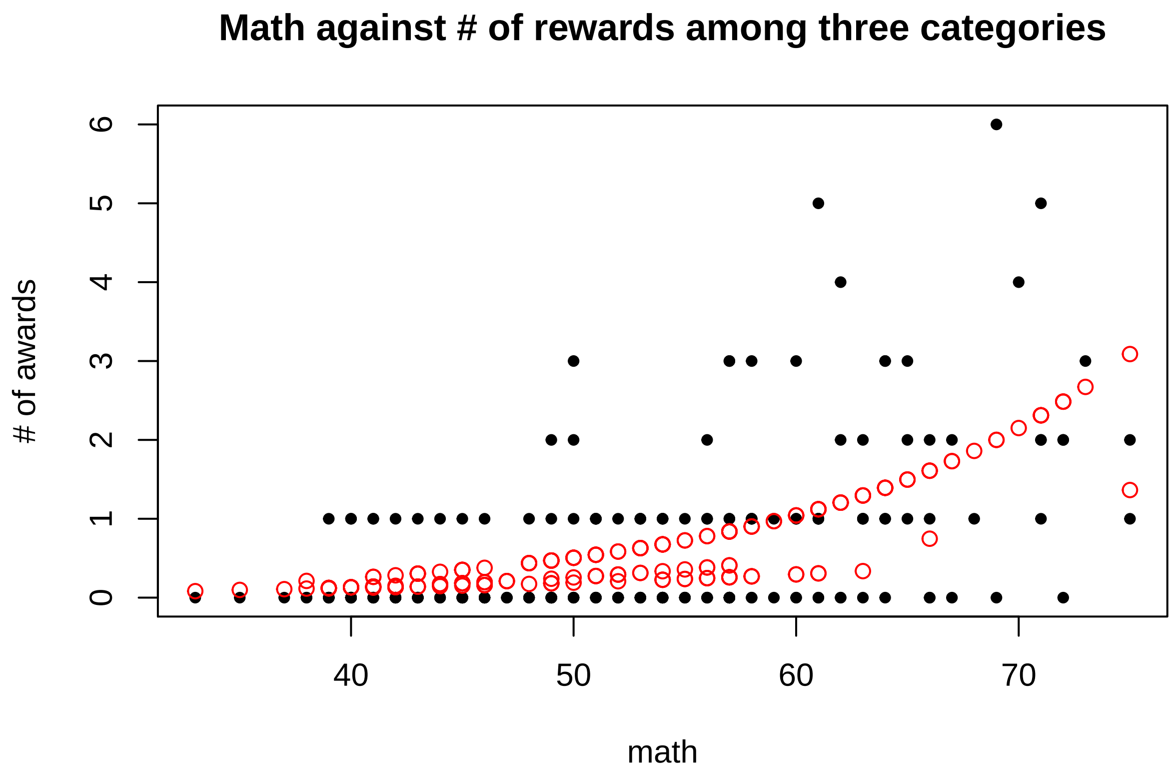

```r
plot(vocational$math,vocational$numawards,col="black",pch=20,xlab="math",ylab = "# of awards",main = "St
points(vocational$math,vocational$m4,col="red")
legend(x=3.5,y=14, legend=c("True", "Predicted"),fill=c("black", "red"))
```



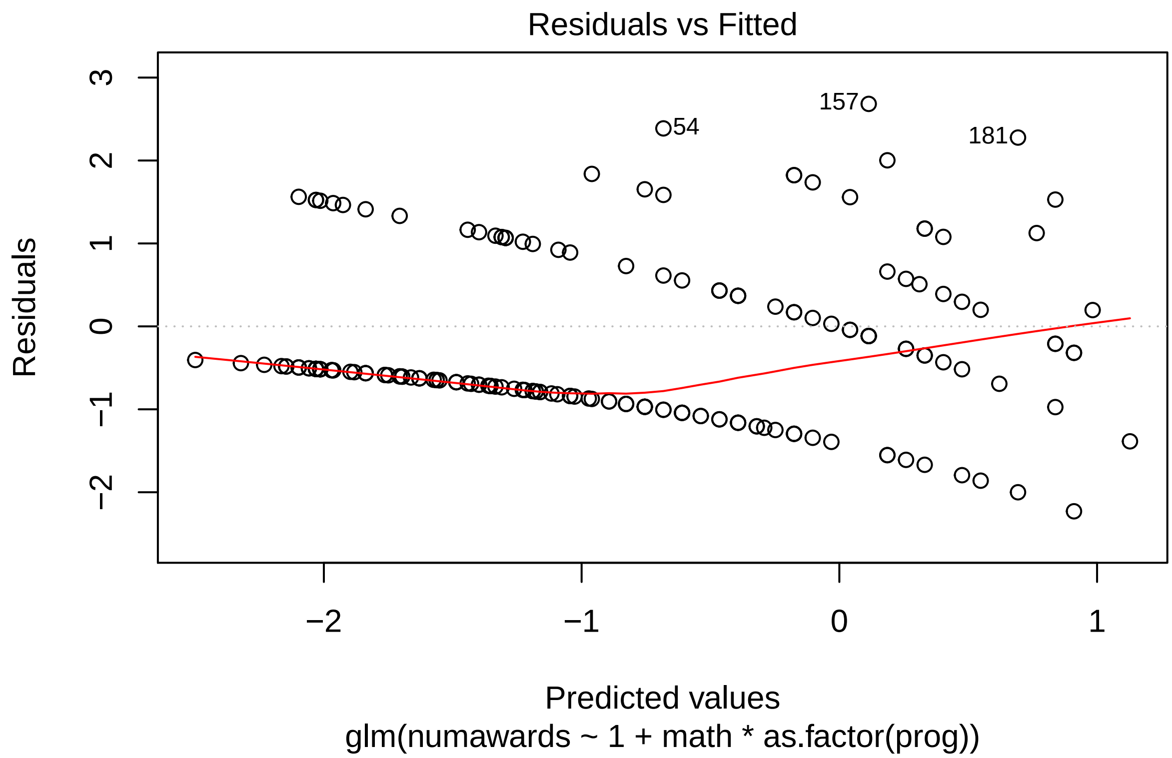

```
plot(general$math,general$numawards,col="black",pch=20,xlab="math",ylab = "# of awards",main = "Students
points(general$math,general$m4,col="red")
legend(x=3.5,y=14, legend=c("True", "Predicted"),fill=c("black", "red"))
```

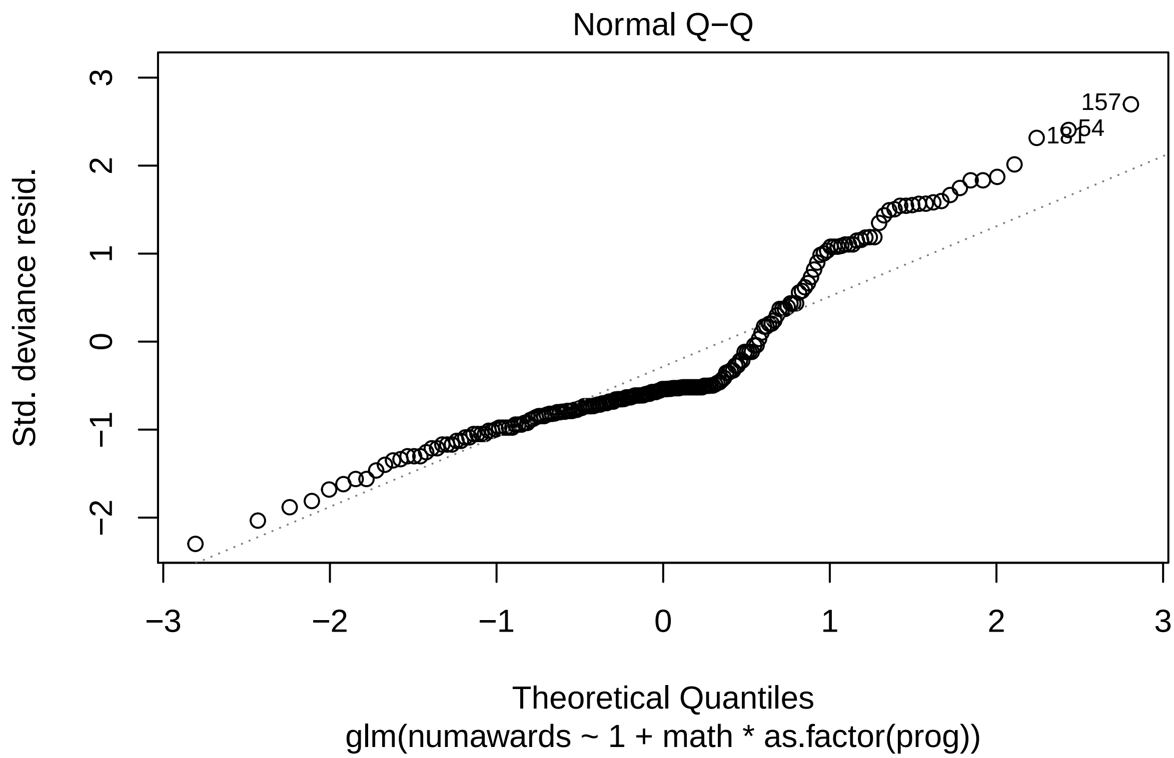


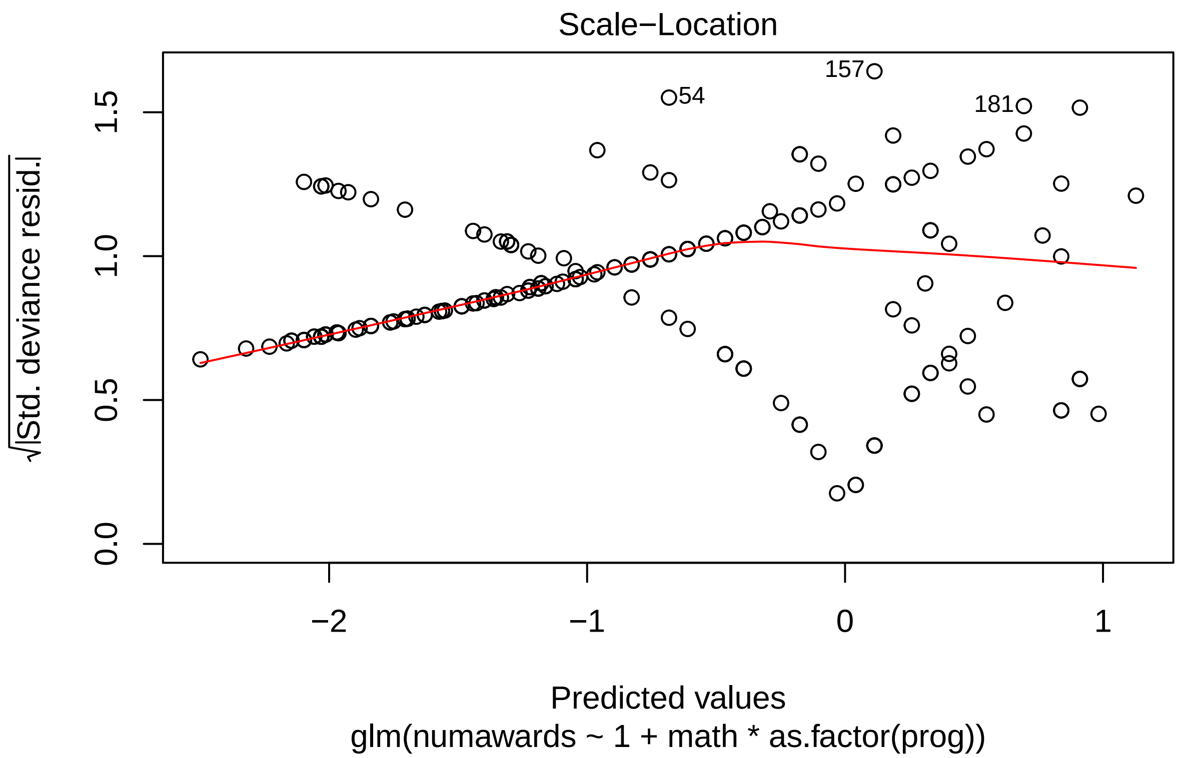

```
plot(awards$math,awards$numawards,col="black",pch=20,xlab="math",ylab = "# of awards",main = "Math agai
points(awards$math,awards$m4,col="red")
legend(x=3.5,y=14, legend=c("True", "Predicted"),fill=c("black", "red"))
```

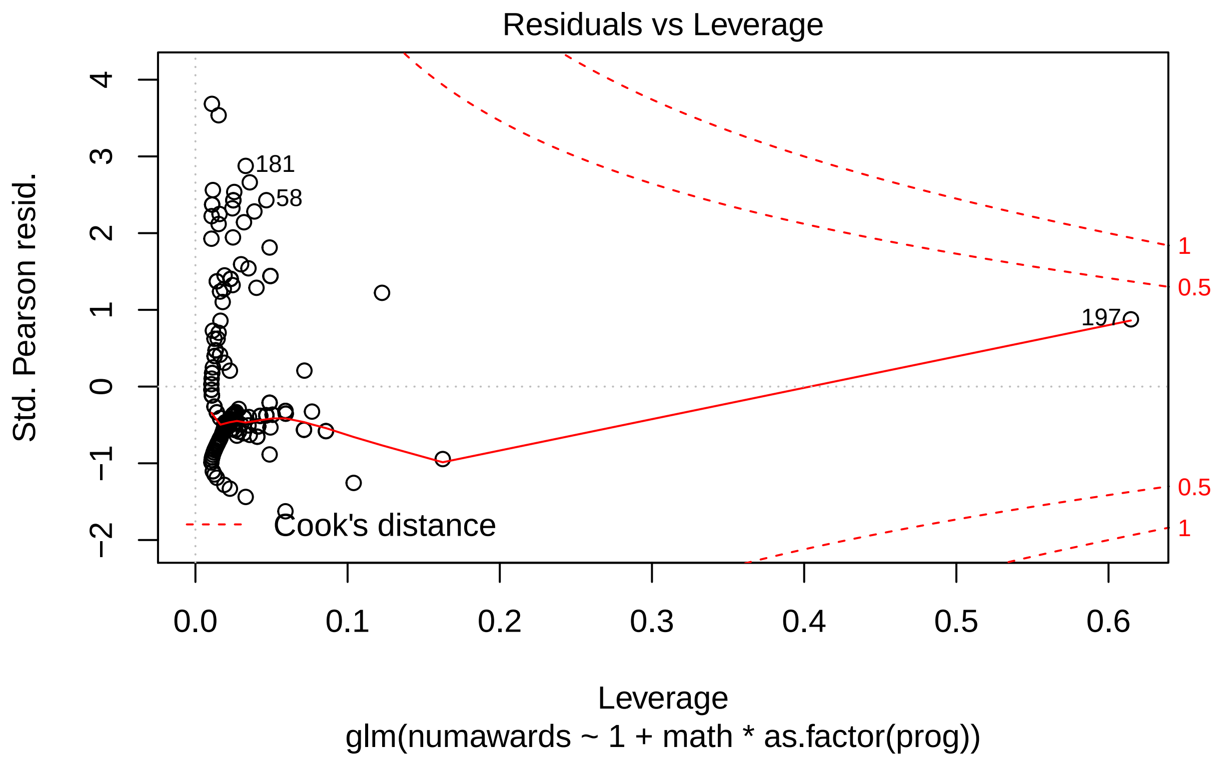

Math against # of rewards among three categories
# of awards
0
1
2
3
4
5
6
40
50
60
70
math

```
plot(m4)
```


Residuals vs Fitted
157
54
181
Residuals
3
2
1
0
−1
−2
−2
−1
0
1
Predicted values
glm(numawards ~ 1 + math * as.factor(prog))


Normal Q−Q
Std. deviance resid.
Theoretical Quantiles
glm(numawards ~ 1 + math * as.factor(prog))
157
54
181
−3
−2
−1
0
1
2
3


Scale−Location
157
54
181
1.5
1.0
0.5
0.0
√|Std. deviance resid.|
−2
−1
0
1
Predicted values
glm(numawards ~ 1 + math * as.factor(prog))



Compare Deviance

```
m1$deviance
```

```
## [1] 204.0213
```

```
m2$deviance
```

```
## [1] 234.46
```

```
m3$deviance
```

```
## [1] 189.4496
```

```
m4$deviance
```

```
## [1] 189.1016
```

Based on the deviance of model 1 to 4, we can conclude that model 4 is the best with smallest deviance 189.1016.