Stats 330 A2

Richard Choi 915726645

Due Date: 3pm Friday 30 August

(i) Create a new section with an appropriate heading so that the markers can easily navigate through your report.

(ii) Include R code with any plots or output.

(iii) Comment briefly means write a few sentences about what you discovered from this process.

(iv) Please remember to handin your hand copy, with signed cover sheet, by the due date.

Compute the goodness of fit statistics for each of these three models and compare them. Comment briefly.

```
MK.df = read.csv("masskill.csv")
model.null <- glm(masskill ~ 1, family = "poisson", offset = log(popn/100), d
ata = MK.df)

model.lin <- glm(masskill ~ I(year - 1982), family = "poisson", offset = log(
popn/100), data = MK.df)

model.quad <- glm(masskill ~ I(year - 1982) + I((year - 1982)^2), family = "p
oisson", offset = log(popn/100), data = MK.df)

# goodness of fit for null model
1 - pchisq(71.466, 36)

## [1] 0.000393921

# goodness of fit for linear model
1 - pchisq(36.443, 35)

## [1] 0.4014187
```

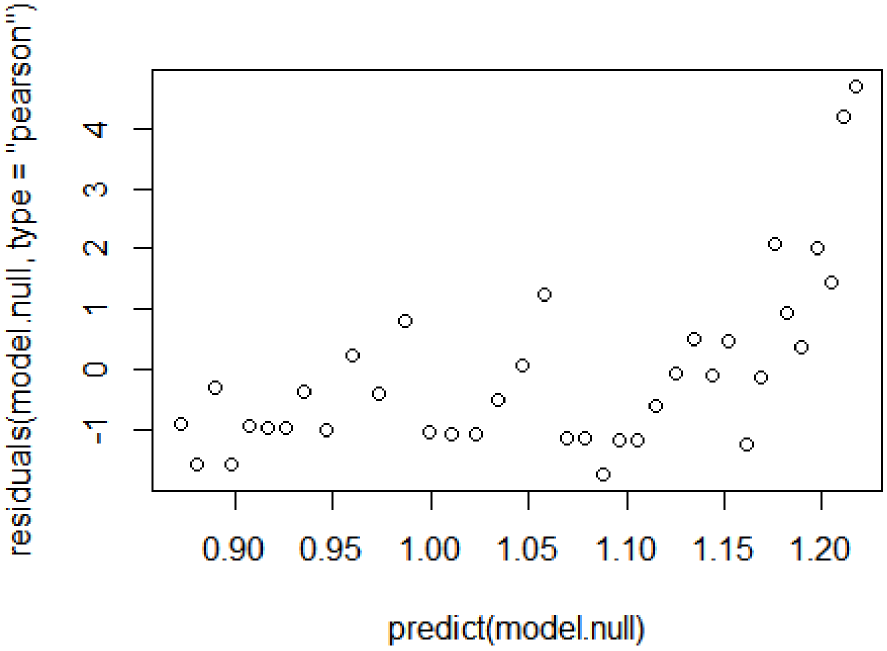
```
# goodness of fit for quadratic model
1 - pchisq(31.142, 34)

## [1] 0.6084442
```

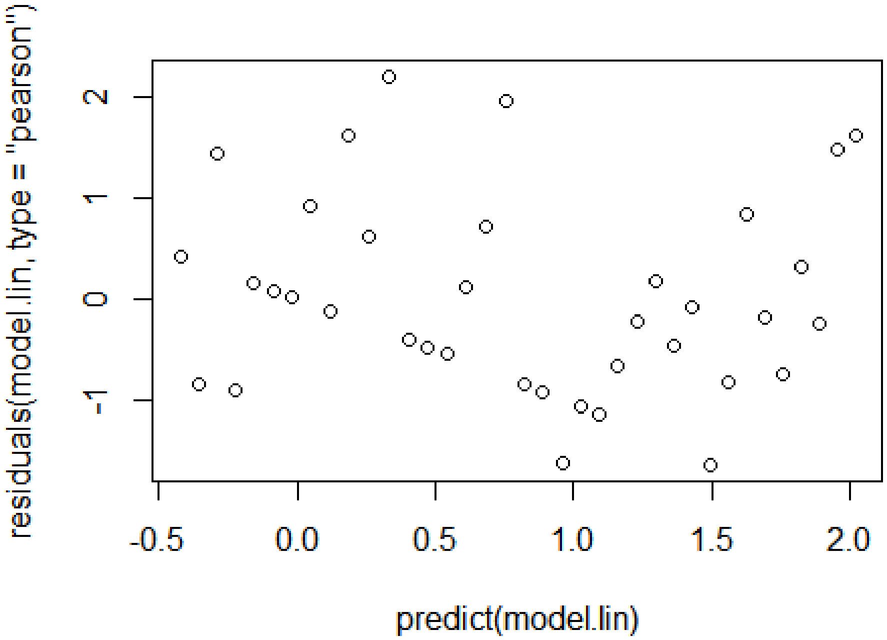
The goodness of fit tells that the null model is overdispersed which indicates that the model is not siuitable whereas, the linear model and quadratic model are fine so we have no concern abou the Poisson model.

Compare the Pearson residual plots for each of these three models. Comment briefly

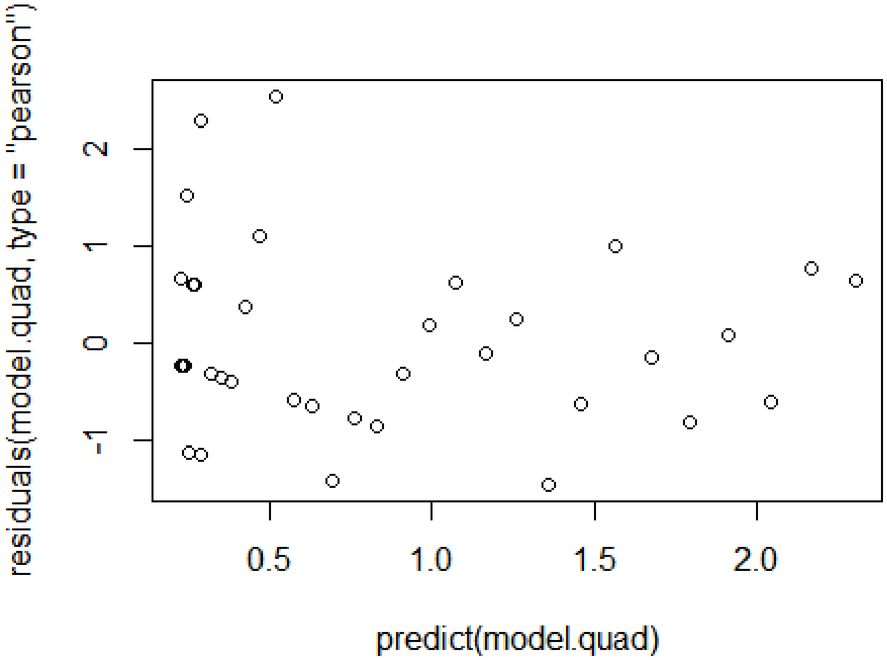
```
plot(predict(model.null), residuals(model.null, type = "pearson"))
```



```
plot(predict(model.lin), residuals(model.lin, type = "pearson"))
```



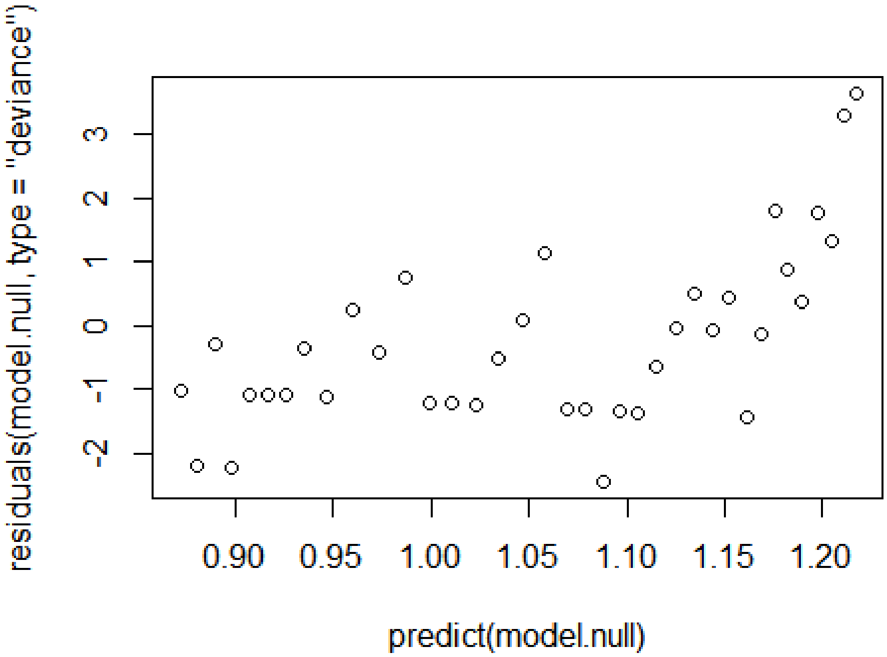
```
plot(predict(model.quad), residuals(model.quad, type = "pearson"))
```



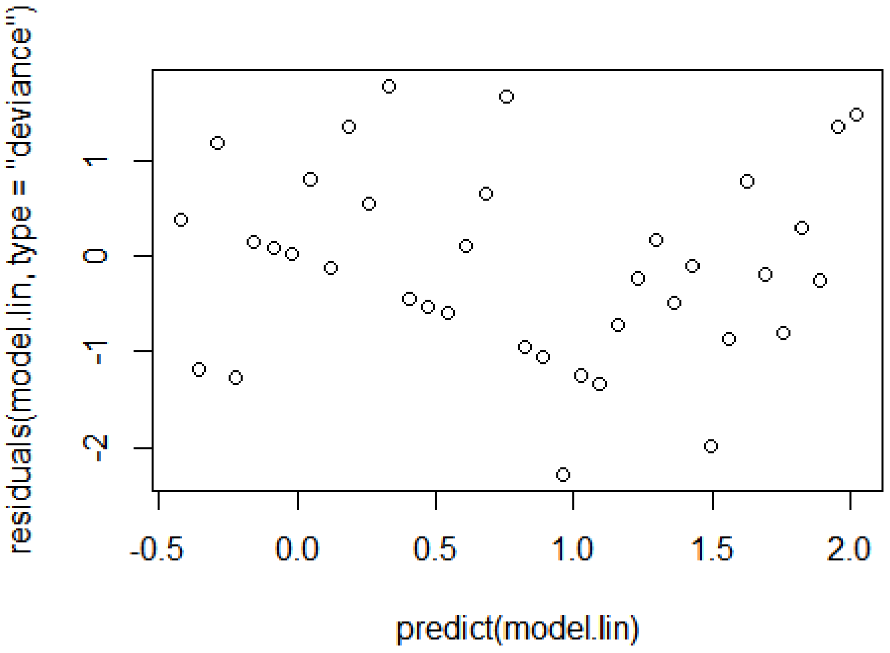
The Pearson residual of null model appears to have its variance increase which indicates that the null model is not appropriate. Whereas, the pearson residual plots of linear model and quadratic model seem to have roughly constant variance and have mean zero across the range of fitted values.

Compare the Deviance residual plots for each of these three models. Comment briefly.

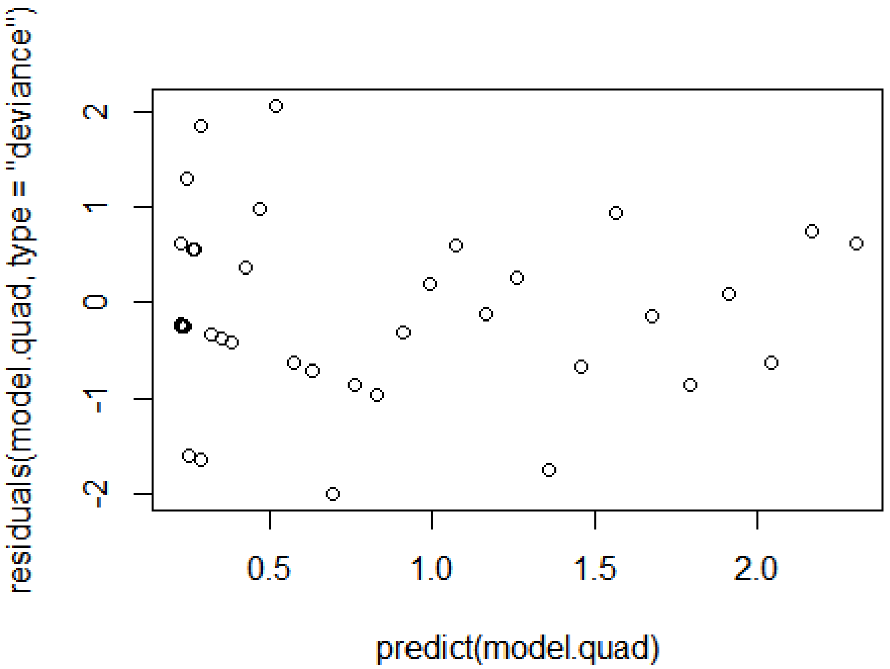
```
plot(predict(model.null), residuals(model.null, type = "deviance"))
```



```
plot(predict(model.lin), residuals(model.lin, type = "deviance"))
```



```
plot(predict(model.quad), residuals(model.quad, type = "deviance"))
```



We notice that the deviance residual plots for three models are similar to the pearson residual plot. We also observe the deviance residual for null model does not have constant variance whereas, the rest have constant variance and have mean zero across the range of fitted values.

These three models are nested - explain why they are and perform a deviance comparison test (ANOVA) of all three models. Comment briefly.

```
anova(model.lin, model.quad, test = "Chisq")

## Analysis of Deviance Table
## 
## Model 1: masskill ~ I(year - 1982)
## Model 2: masskill ~ I(year - 1982) + I((year - 1982)^2)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1        35     36.443                       
## 2        34     31.142  1   5.3004  0.02132 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

anova(model.quad, model.null, test = "Chisq")

## Analysis of Deviance Table
## 
## Model 1: masskill ~ I(year - 1982) + I((year - 1982)^2)
## Model 2: masskill ~ 1
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)    
## 1        34     31.142                          
## 2        36     71.466 -2  -40.323 1.753e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

anova(model.lin, model.null, test = "Chisq")

## Analysis of Deviance Table
## 
## Model 1: masskill ~ I(year - 1982)
## Model 2: masskill ~ 1
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)    
## 1        35     36.443                          
## 2        36     71.466 -1  -35.023 3.258e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

These three models are nested because the linear model and null model are submodel of quadratic model. This is because if we set parameter values of quadratic model to zero then we can end up with either linear model and null model. We have a strong evidence to suggest that we need need at least one of the qudratic term in the full model (p - value =

0.02132). Moreover, we found a very strong evidnece that we need at least linear term in the full model (p - value = 3.258e-09). We also found a strong evidence that we need at least one linear term or quadratic term in the full model (p - value = 1.753e-09).

Compute the AIC, AIcc, and BIC statistics for these three models. Comment briefly.

```
library(MuMIn)

## Warning: package 'MuMIn' was built under R version 3.5.3

AIC(model.null, model.lin, model.quad)

##            df      AIC
## model.null  1 167.3777
## model.lin   2 134.3547
## model.quad  3 131.0543

AICc(model.null, model.lin, model.quad)

##            df     AICc
## model.null  1 167.4920
## model.lin   2 134.7076
## model.quad  3 131.7816

BIC(model.null, model.lin, model.quad)

##            df      BIC
## model.null  1 168.9886
## model.lin   2 137.5765
## model.quad  3 135.8871
```

We found that the null model, linear model, and quadratic model are ranked highest from lowest respectively by AIC and AICc, and BIC measurement.

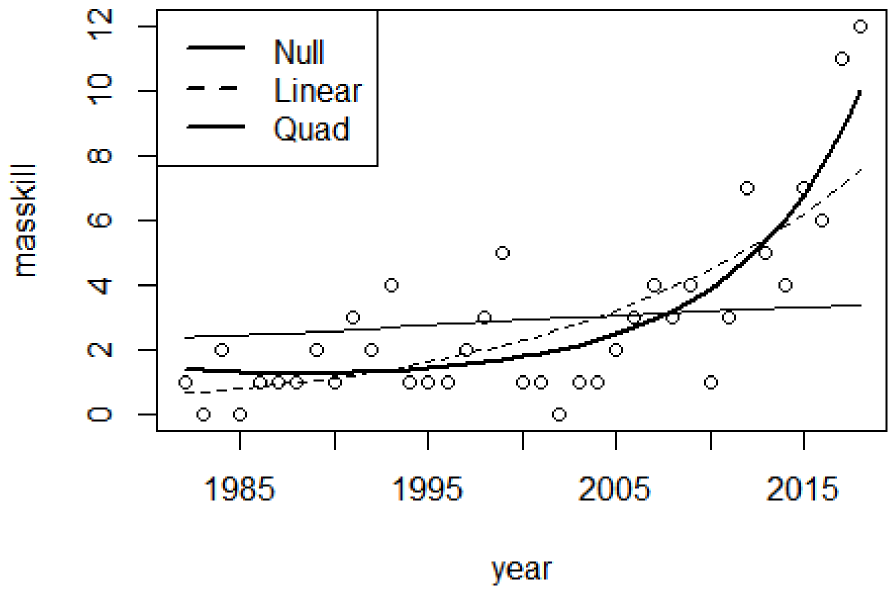
Plot these data as you did assignment. Overlay the predicted expected mass killings for each of these three models. (Hints: the R functions predict, lines will be useful here.) In particular, does the above analysis change any of the conclusions you made at the end of Assignment 1? Comment briefly.

```
plot(masskill ~ year, data = MK.df)
null_pred <- predict(model.null, MK.df, se.fit = TRUE)
lines(x=1982:2018, exp(null_pred$fit))

lin_pred <- predict(model.lin, MK.df, se.fit = TRUE)
lines(x=1982:2018, exp(lin_pred$fit), lty = 2)
```

```
quad_pred <- predict(model.quad, MK.df, se.fit = TRUE)
lines(x=1982:2018, exp(quad_pred$fit), lwd = 2)

legend("topleft", legend = c("Null", "Linear", "Quad"), lty = 1:2, lwd = 2)
```



We can see that the null model is increasing constantly due to the exposure variable as there is more population over time. The linear model is increasing linearly as expected and the quadratic model is increasing with a curvature as expected as well. The conclusion I made in the previous assignment was that the linaer model was appropriate and the above analysis helped me change my conclusion. This is because AIC, AICc, and BIC measurements indicate that the quadratic model is the best model out of them. This means that the quadratic model is the best candidate, it has lower deviance, higher log-likelihood and have small number of parameters in comparison out of all the models. The Pearson and deviance residual plots of quadratic model was patternless and had a constant variance. We also found a strong evidence that at least one quadratic term is needed in the final model.

They started by fititng the following model to start with and then got distracted. Your job is to complete the task they started by seeing if you can create a simpler model than the rather complex one above.

```
library(statmod)
```

```
## Warning: package 'statmod' was built under R version 3.5.3
```

```
Stats20x.df = read.table("STATS20x.txt", header = T)
full.model <- glm(Pass ~ Test + I(Test^2) + Assign + I(Assign^2) + Attend + R
epeat, family = "binomial", data = Stats20x.df)
```

```
options(na.action = "na.fail")
all.fits <- dredge(full.model)
```

```
## Fixed term is "(Intercept)"
```

```
head(all.fits)
```

```
## Global model call: glm(formula = Pass ~ Test + I(Test^2) + Assign + I(Assi
gn^2) + 
##     Attend + Repeat, family = "binomial", data = Stats20x.df)
## ---
## Model selection table 
##    (Intrc)  Assgn  Assgn^2 Attnd   Test   Test^2 df  logLik AICc delta
## 22  -16.92 0.6587              + 0.8447           4 -24.270 56.8  0.00
## 38  -13.12 0.6693              +         0.04287  4 -24.764 57.8  0.99
## 18  -16.05 0.7002                0.7869           3 -25.861 57.9  1.07
## 23  -12.91         0.02750     + 0.8117           4 -24.828 57.9  1.12
## 54  -19.59 0.6529              + 1.4370 -0.03030  5 -24.168 58.8  1.94
## 24  -18.79 0.9659 -0.01259     + 0.8539           5 -24.226 58.9  2.06
##    weight
## 22  0.285
## 38  0.174
## 18  0.167
## 23  0.163
## 54  0.108
## 24  0.102
## Models ranked by AICc(x)
```

```
good.model <- get.models(all.fits, 1)[[1]]
good.model
```

```
## 
## Call:  glm(formula = Pass ~ Assign + Attend + Test + 1, family = "binomial
", 
##     data = Stats20x.df)
## 
## Coefficients:
## (Intercept)       Assign    AttendYes        Test  
##    -16.9249       0.6587       1.3914       0.8447  
## 
## Degrees of Freedom: 145 Total (i.e. Null);  142 Residual
## Null Deviance:       178.7 
## Residual Deviance: 48.54     AIC: 56.54
```

```
100 * (exp(confint(good.model)) - 1)
```

```
## Waiting for profiling to be done...
##                   2.5 %     97.5 %
## (Intercept) -100.00000  -99.99853
## Assign        46.84049  183.23034
## AttendYes    -12.69050 2090.78869
## Test          65.51229  270.54379
```

Discuss how you arrived at your final model

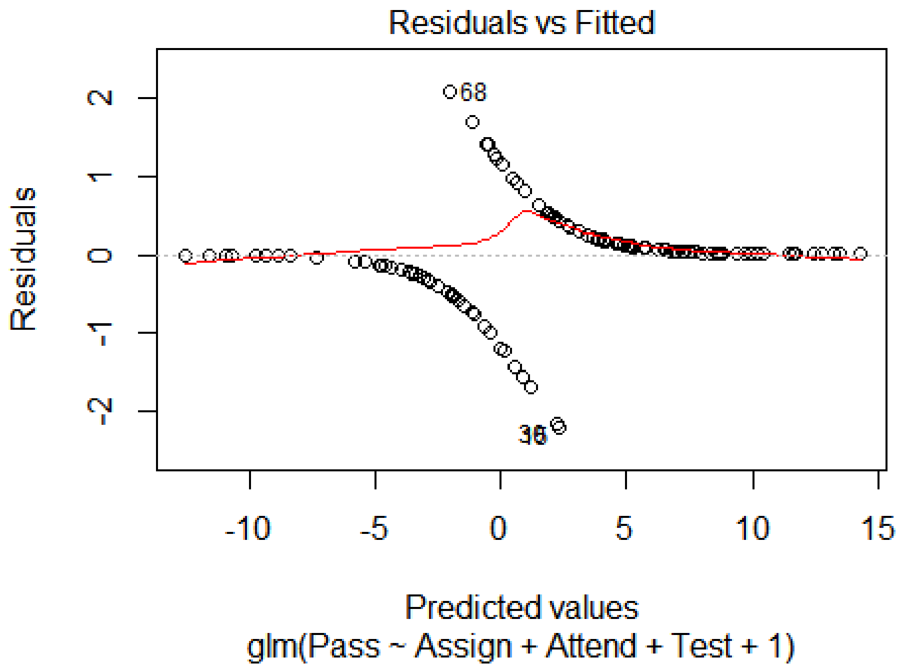
a) I've arrived at my final model by inputting the full model so that the function will fit all possible submodels and compare the models' AICc. Hence, the model without any quadratic term, repeat variable, and an addition of y-intercept of 1 came out with the lowest value of AICc. This means that the final model is the best model candidate, it has low deviance, high log-likelihood, but also a small number of parameters. The sample sizes are large, so we have no conern of using AICc. Athough, the attend variable was not statistically significant I decided to keep the variable as it only means we have weak evidence against the null hypothesis, doesn't mean it has no effect on students passing. Moreover, I also believe that students attending the lecture has effect on students passing the course so I retained it.

Make sure that your model assumptions are satisfied.

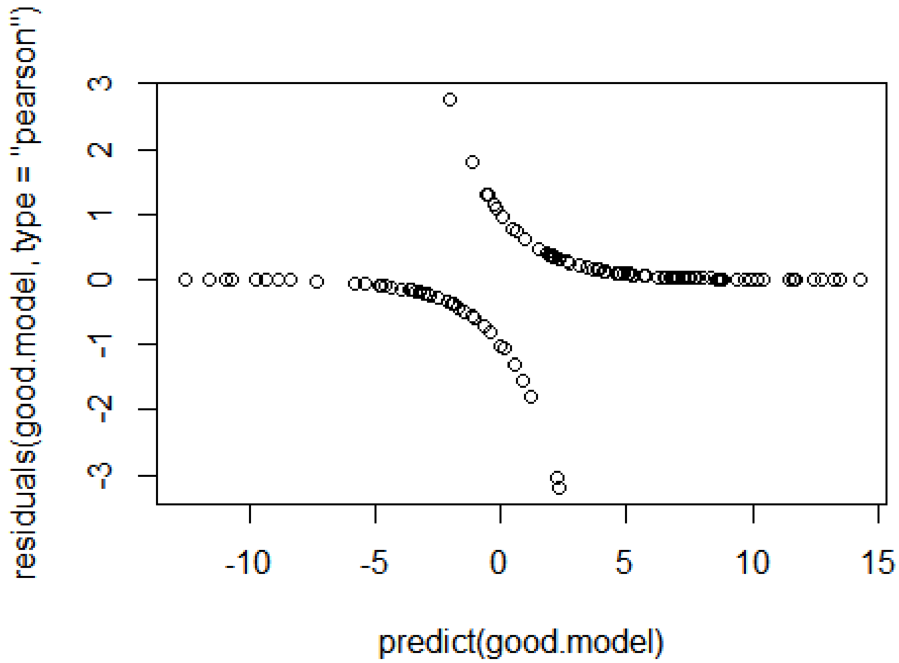
```
summary(good.model)

## 
## Call:
## glm(formula = Pass ~ Assign + Attend + Test + 1, family = "binomial", 
##     data = Stats20x.df)
## 
## Deviance Residuals: 
##      Min        1Q    Median        3Q       Max  
## -2.19882  -0.02083   0.02928   0.17806   2.07703  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -16.9249     3.5242  -4.803 1.57e-06 ***
## Assign        0.6587     0.1643   4.008 6.11e-05 ***
## AttendYes     1.3914     0.8032   1.732   0.0832 .  
## Test          0.8447     0.2018   4.186 2.84e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 178.71  on 145  degrees of freedom
## Residual deviance:  48.54  on 142  degrees of freedom
## AIC: 56.54
## 
## Number of Fisher Scoring iterations: 8
```

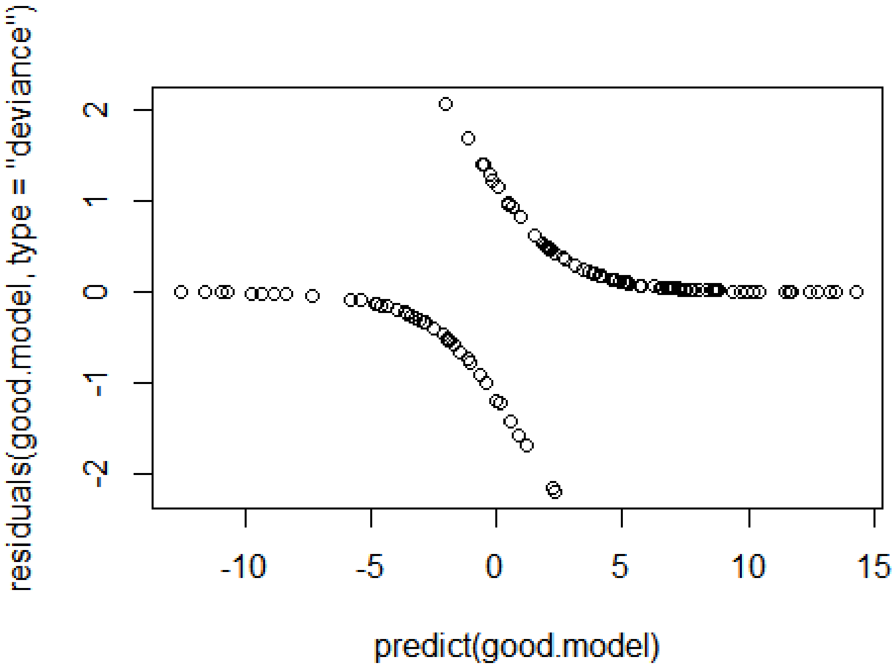
```r
plot(good.model, which = 1)
```



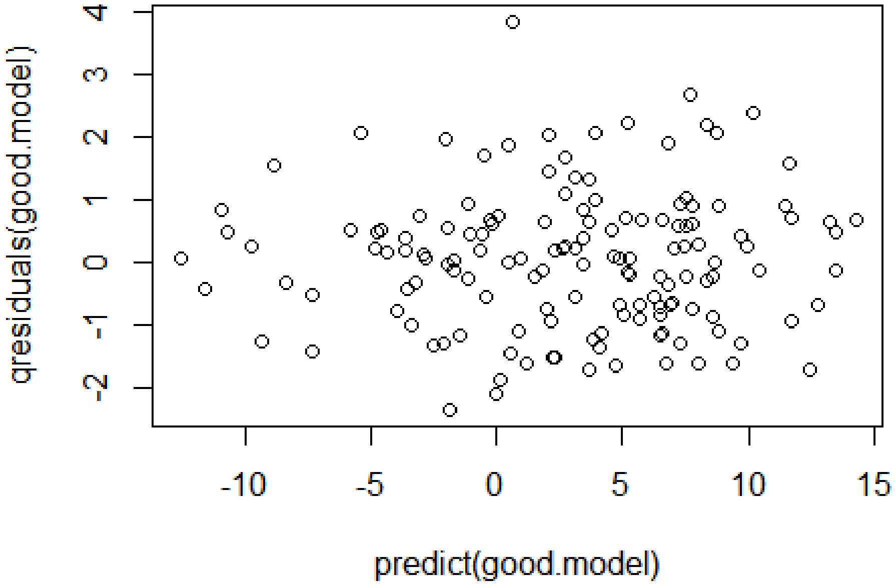
```r
plot(predict(good.model), residuals(good.model, type = "pearson"))
```



```r
plot(predict(good.model), residuals(good.model, type = "deviance"))
```



```r
plot(predict(good.model), qresiduals(good.model))
```



As the response variable is a categorical variable for students passing stats 20x or not, we have therefore fitted a generalised linear model with a binomial response distribution. The data set is a random sample of students so we can safely assume that the data set is independent. Since, the data is binary, Chisquared test for residual deviance was not used to test for goodness of fit purpose. We can observe that the variability increases around 0 as we are fitting binomial data from raw residual as expected from binomial distribution. Moreover, the Pearson residual and deviance residual showed similar shape with the raw residual plot which wasn't helpful. Therefore, qresidual plot was used to test if the model is correct. The randomised quantile residual showed patternless and approximately normal which indicates that the model is correct.

Write a brief executive summary about your final model.

c) We aimed to create a model that predicts whether or not a student passes 20x based on their performanc and behaviour during the year (and previously). We found a strong evidence that there is a positive relationship for assignment mark (p-value = 6.11e-05) and test mark (p-value = 2.84e-05) with students passing the 20x course. However, we found there is a weak evidence that there is a positive relationship between students attending the class and the students passing the 20x course (p-value = 0.0832). Also, we estimate that every 1 mark increase in assignment corresponds to an increase in odds of a student passing 20x course by between 47% and 183%. We estimate for every 1 mark increase in test leads to an increase in odds of a student passing 20x course by between 66% and 271%.