HW4 Peer Assessment

Mon Oct 24 23:24:02 2022 EDT

Background

The owner of a company would like to be able to predict whether employees will stay with the company or leave.

Data Description

The data contains information about various characteristics of employees. Please note that the dataset has been updated to account for repetitions, which is needed for Goodness of Fit Assessment. See below for the description of these characteristics.

1. **Age.Group**: 1-9 (1 corresponds to teen, 2 corresponds to twenties, etc.)
2. **Gender**: 1 if male, 0 if female
3. **Tenure**: Number of years with the company
4. **Num.Of.Products**: Number of products owned
5. **Is.Active.Member**: 1 if active member, 0 if inactive member
6. **Staying**: Fraction of employees that stayed with the company for a given set of predicting variables.

Setup

You can import the data and set up the problem with the following R code:

```
# Import the data
rawdata = read.csv("hw4_data.csv", header=TRUE, fileEncoding="UTF-8-BOM")

# Create variable Staying
rawdata$Staying = rawdata$Stay/rawdata$Employees

# Set variables as categoricals
rawdata$Num.Of.Products<-as.factor(rawdata$Num.Of.Products)
rawdata$Age.Group<-as.factor(rawdata$Age.Group)
rawdata$Gender<-as.factor(rawdata$Gender)
rawdata$Is.Active.Member<-as.factor(rawdata$Is.Active.Member)

# Print head of rawdata
head(rawdata)
```

	Age.Group <fct>	Gender <fct>	Tenure <int>	Num.Of.Products <fct>	Is.Active.Member <fct>	Stay <int>	Employees <int>	Staying <dbl>
1	2	1	3	1	0	5	11	0.4545455
2	2	1	4	1	0	5	10	0.5000000
3	2	1	4	1	1	2	13	0.1538462
4	2	0	7	1	0	3	10	0.3000000
5	2	1	7	1	0	2	14	0.1428571
6	2	0	4	2	0	4	12	0.3333333

6 rows

Note: For all of the following questions, treat variables **Tenure** and **Staying** as quantitative variables and **Age.Group**, **Gender**, **Num.Of.Products**, and **Is.Active.Member** as categorical variables. Categorical variables have already been converted to factors in the starter code.

Question 1: Fitting a Model - 9 pts

Fit a logistic regression model using *Staying* as the response variable with *Num.Of.Products* as the predictor and logit as the link function. Ensure to include the weights parameter for specifying the number of trials. Call it **model1**. Note that *Num.Of.Products* should be treated as a categorical variable.

a. 3 pts - Display the summary of model1. What are the model parameters and estimates?

```
model1 <- glm(Staying~Num.Of.Products, data=rawdata, weights=Employees, family = binomial(link="l
ogit"))
summary(model1)
```

```
##
## Call:
## glm(formula = Staying ~ Num.Of.Products, family = binomial(link = "logit"),
##     data = rawdata, weights = Employees)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.2827  -1.4676  -0.1022   1.4490   4.7231
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        0.37886    0.04743   7.988 1.37e-15 ***
## Num.Of.Products2 -1.76683    0.10313 -17.132  < 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: 981.04  on 157  degrees of freedom
## Residual deviance: 632.04  on 156  degrees of freedom
## AIC: 1056.8
##
## Number of Fisher Scoring iterations: 4
```

Answer 1a:

The model parameters are B0(intercept) and B1(Num.Of.Products2)
B0 = 0.37886
B1 = -1.76683

b. 3 pts - Write down the equation for the Odds of Staying.

Answer 1b:

p(staying)/1-p(staying) = e^(B0 + B1(Num.Of.Products2)) = e^(0.37886+-1.76683(Num.Of.Products2))

c. 3 pts - Provide a meaningful interpretation for the estimated coefficient for *Num.Of.Products2* with respect to the log-odds of staying and the odds of staying.

Answer 1c:

For every unit of increase in Num.Of.Products2 the log odds Employees staying decreases by -1.76683.

For every unit of increase in Num.Of.Products2 the odds of Employees staying increases by e^-1.76683.

Question 2: Inference - 9 pts

a. 3 pts - Using model1, find a 90% confidence interval for the coefficient for *Num.Of.Products2*.

```
confint(model1, 'Num.Of.Products2', level = 0.90)
```

```
## Waiting for profiling to be done...
```

```
##       5 %      95 %
## -1.938361 -1.598965
```

Answer 2a:

The confidence interval at 90% is [-1.938361,-1.598965]

b. 3 pts - Is model1 significant overall at the 0.01 significance level?

```
1-pchisq((model1$null.deviance-model1$deviance), (model1$df.null-model1$df.residual))
```

```
## [1] 0
```

Answer 2b:

The results p value obtained from the chi-sq test is returning 0 meaning that it is less than 0.01 significance level, so we reject the null hypothesis and therefore model1 is significant overall.

c. 3 pts - Which regression coefficients are significantly nonzero at the 0.01 significance level? Which are significantly negative? Why?

Answer 2c:

Both the p-value of the intercept and Num.Of.Products2 are less than alpha=0.01, therefore significantly nonzero. Num.Of.Products2 is significantly negative because it has a negative coefficient.

Question 3: Goodness of fit - 10 pts

a. 3.5 pts - Perform goodness-of-fit hypothesis tests using both Deviance and Pearson residuals. What do you conclude? Explain the differences, if any, between these findings and what you found in Question 2b.

```
#Deviance test for GOF
c(deviance(model1), 1-pchisq(deviance(model1), 156))
```

```
## [1] 632.04   0.00
```

```
#Pearson test for GOF
model1.pears = residuals(model1,type="pearson")
pearson.tvalue = sum(model1.pears^2)
c(pearson.tvalue, 1-pchisq(pearson.tvalue,156))
```

```
## [1] 562.1763   0.0000
```

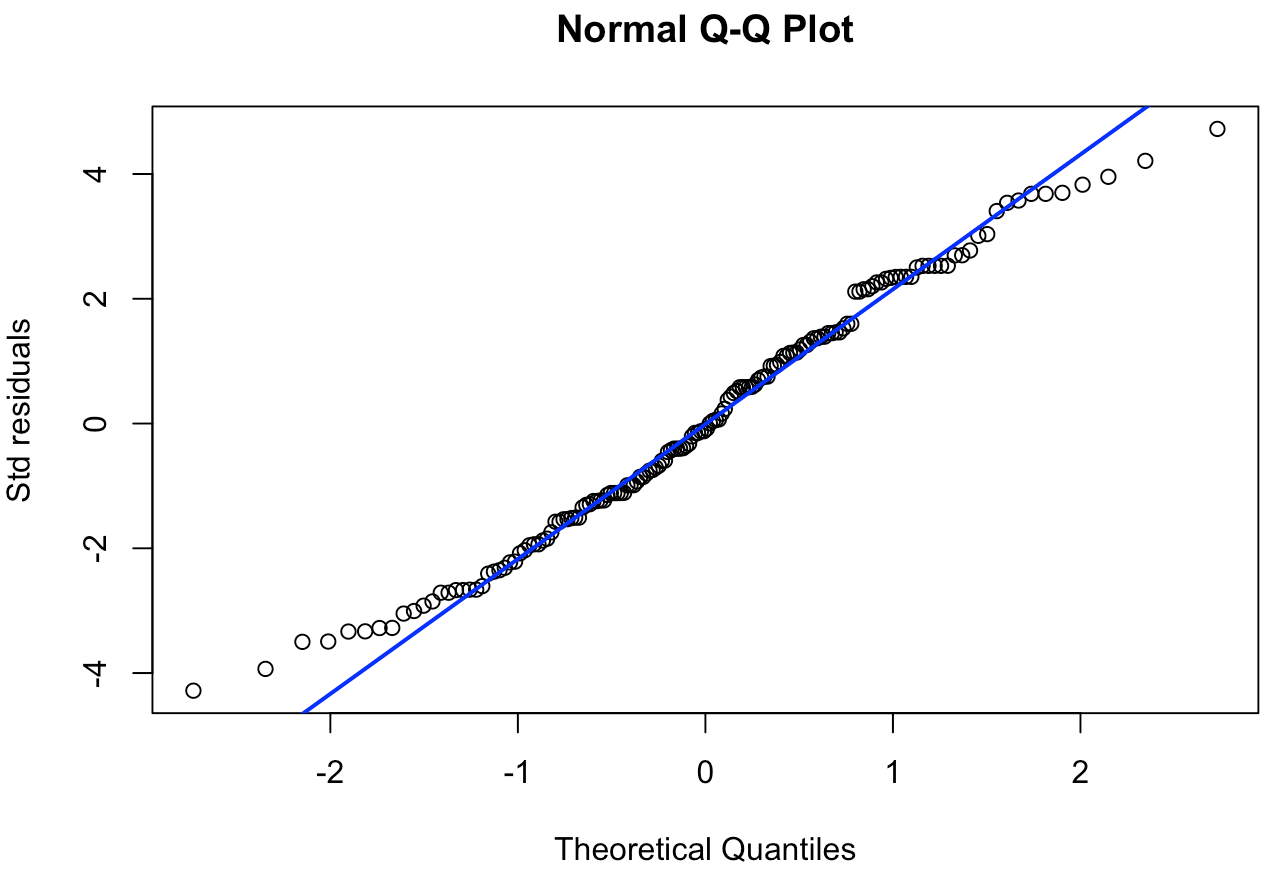
Answer 3a:

The p-value for both the deviance and pearson test is 0. This means that we reject the null hypothesis that the model is a good fit. So the model is not a good fit, which does not agree with findings from 2b. The results from 2b show that goodness of fit does not necessarily impact the predictive power of a model.

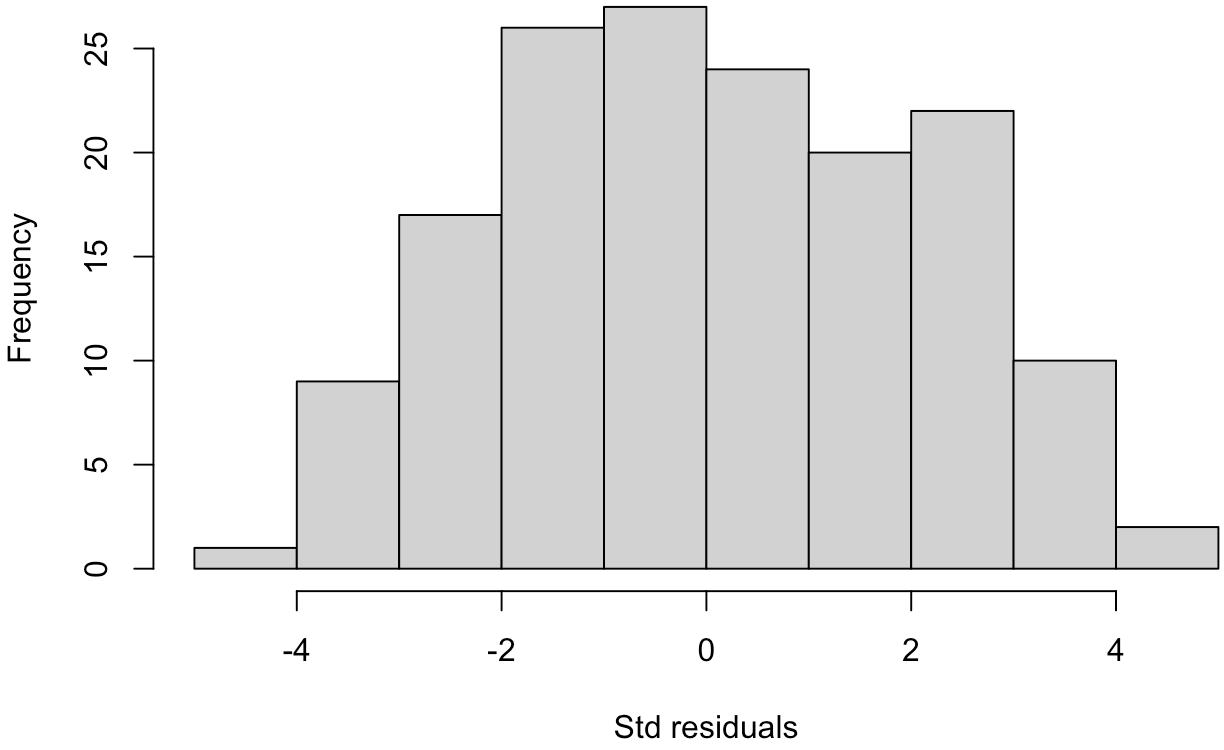
b. 3.5 pts - Evaluate whether the deviance residuals are normally distributed by producing a QQ plot and histogram of the deviance residuals. What assessments can you make about the goodness of fit of **model1** based on these plots?

```
model1.res <- resid(model1,type='deviance')

#QQplot
qqnorm(model1.res, ylab="Std residuals")
qqline(model1.res,col="blue",lwd=2)
```




```
#Histgram
hist(model1.res,10,xlab="Std residuals", main="")
```



Answer 3b:

The QQplot depicts skew on both tails and the histogram does not have a uniform distribution. The normality assumption does not hold based on these plots.

c. 3 pts - Calculate the estimated dispersion parameter for this model. Is this an overdispersed model?

```
model1$deviance/model1$df.residual
```

```
## [1] 4.051539
```

Answer 3c:

The estimated dispersion is 4.05 which is greater than 2 which means that this model is overdispersed.

Question 4: Fitting the full model- 23 pts

Fit a logistic regression model using *Staying* as the response variable with *Age.Group*, *Gender*, *Tenure*, *Num.Of.Products*, and *Is.Active.Member* as the predictors and logit as the link function. Ensure to include the weights parameter for specifying the number of trials. Call it **model2**. Note that Age.Group, Gender, Num.Of.Products, and Is.Active.Member should be treated as categorical variables.

```
model2 <- glm(Staying ~ Age.Group+Gender+Tenure+Num.Of.Products+Is.Active.Member, data=rawdata, w
eight=Employees, family = binomial(link = "logit"))
summary(model2)
```

```
##
## Call:
## glm(formula = Staying ~ Age.Group + Gender + Tenure + Num.Of.Products +
##     Is.Active.Member, family = binomial(link = "logit"), data = rawdata,
##     weights = Employees)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -3.10929  -0.76949  -0.07324   0.74079   3.06551
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -0.109572   0.282617  -0.388    0.698
## Age.Group3         0.384480   0.267984   1.435    0.151
## Age.Group4         1.734115   0.270384   6.414 1.42e-10 ***
## Age.Group5         2.955578   0.337727   8.751  < 2e-16 ***
## Gender1           -0.572069   0.093776  -6.100 1.06e-09 ***
## Tenure            -0.003319   0.016569  -0.200    0.841
## Num.Of.Products2  -1.410946   0.112000 -12.598  < 2e-16 ***
## Is.Active.Member1 -0.850280   0.095829  -8.873  < 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: 981.04  on 157  degrees of freedom
## Residual deviance: 162.35  on 150  degrees of freedom
## AIC: 599.07
##
## Number of Fisher Scoring iterations: 4
```

a. 3 pts - Write down the equation for the probability of staying.

Answer 4a:

p(staying)/1-p(staying) = e^(-0.109572 + 0.384480(Age.Group3) + 1.734115(Age.Group4) + 2.955578(Age.Group5) + -0.572069(Gender1) + -0.003319(Tenure) + -1.410946(Num.Of.Products2) + -0.850280(Is.Active.Member1)

b. 3 pts - Provide a meaningful interpretation for the estimated coefficients of *Tenure* and *Is.Active.Member1* with respect to the odds of staying.

Answer 4b:

For every unit of increase in Tenure the odds Employees staying increases by e^-0.003319 if all other factors remain constant.

For every unit of increase in Is.Active.Member1 the odds Employees staying increase by e^-0.850280 if all other factors remain constant.

c. 3 pts - Is *Is.Active.Member1* statistically significant given the other variables in model2 at the 0.01 significance level?

Answer 4c:

Yes, it is significant as the p value is less than 0.01.

d. 10 pts - Has your goodness of fit been affected? Follow the instructions to repeat the tests, plots, and dispersion parameter calculation you performed in Question 3 with **model2**.

(d-1) Perform goodness-of-fit hypothesis tests using both Deviance and Pearson residuals. What do you conclude?

```
#Deviance test for GOF
c(deviance(model2), 1-pchisq(deviance(model2), 150))
```

```
## [1] 162.3494910   0.2319118
```

```
#Pearson test for GOF
model2.pears = residuals(model2,type="pearson")
pearson2.tvalue = sum(model2.pears^2)
c(pearson2.tvalue, 1-pchisq(pearson2.tvalue,150))
```

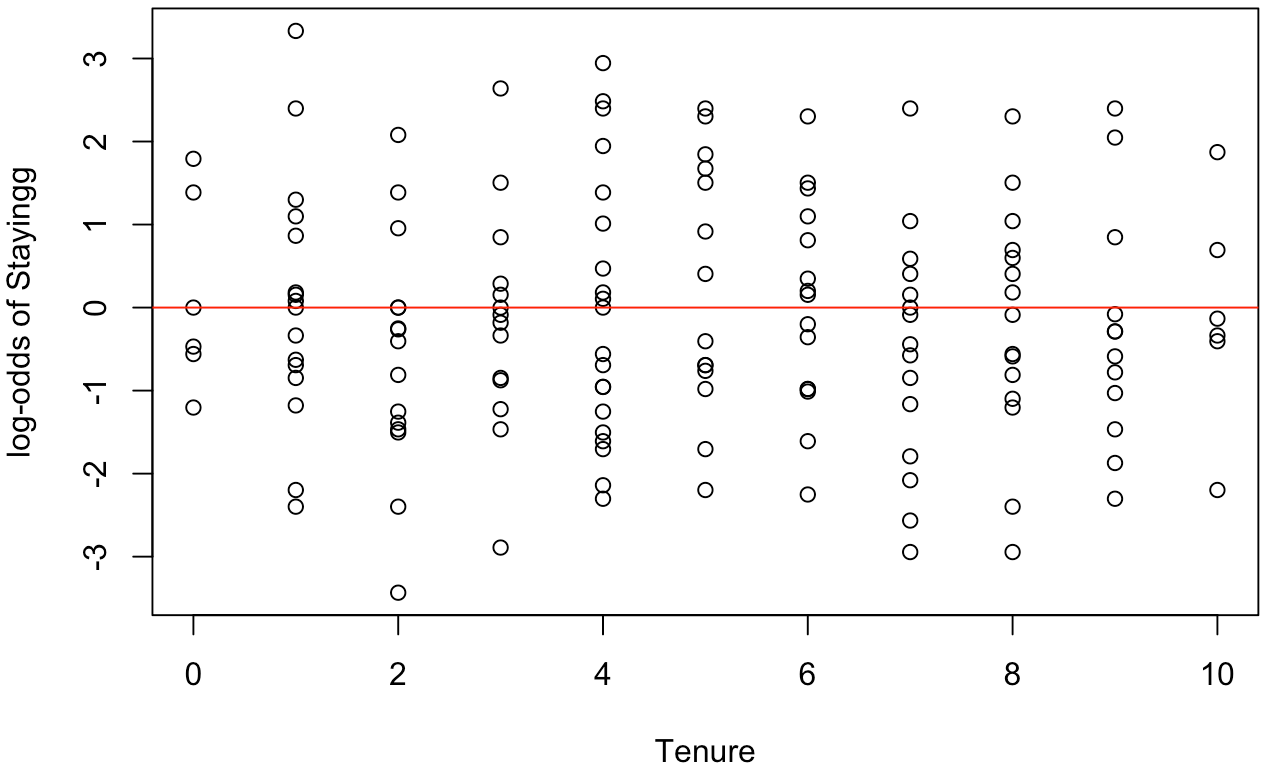
```
## [1] 154.1554225   0.3912174
```

Answer 4d-1:

The p value from the deviance test gives us 0.2319118 and the p value from the pearson test gives us 0.3912174. These values show that the goodness of fit has been affected as they are different from the values of model1. As these values are greater than 0.01, the null hypothesis that the model is a good fit remains.

(d-2) Evaluate the linearity assumption of **model2** by plotting the log-odds of Staying vs. **Tenure**. What do you conclude?

```
Tenure = rawdata$Tenure
Staying = rawdata$Staying
plot(Tenure, log(Staying/(1-Staying)), ylab="log-odds of Stayingg", xlab="Tenure")
abline(0, 0, col="red")
```



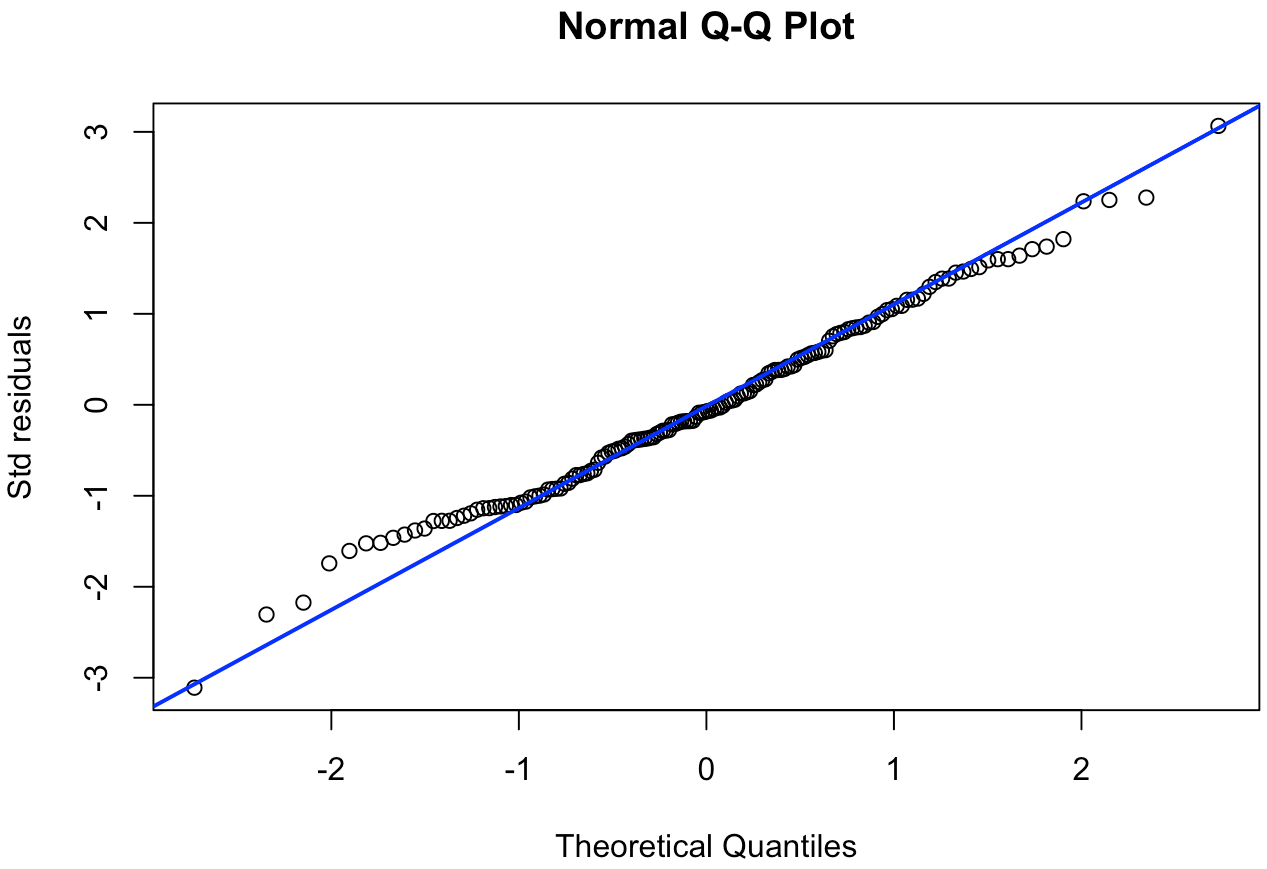
Answer 4d-2:

Based on the logs-odds of staying vs Tenure plot we can see that there little to no relationship between tenure and employees staying.

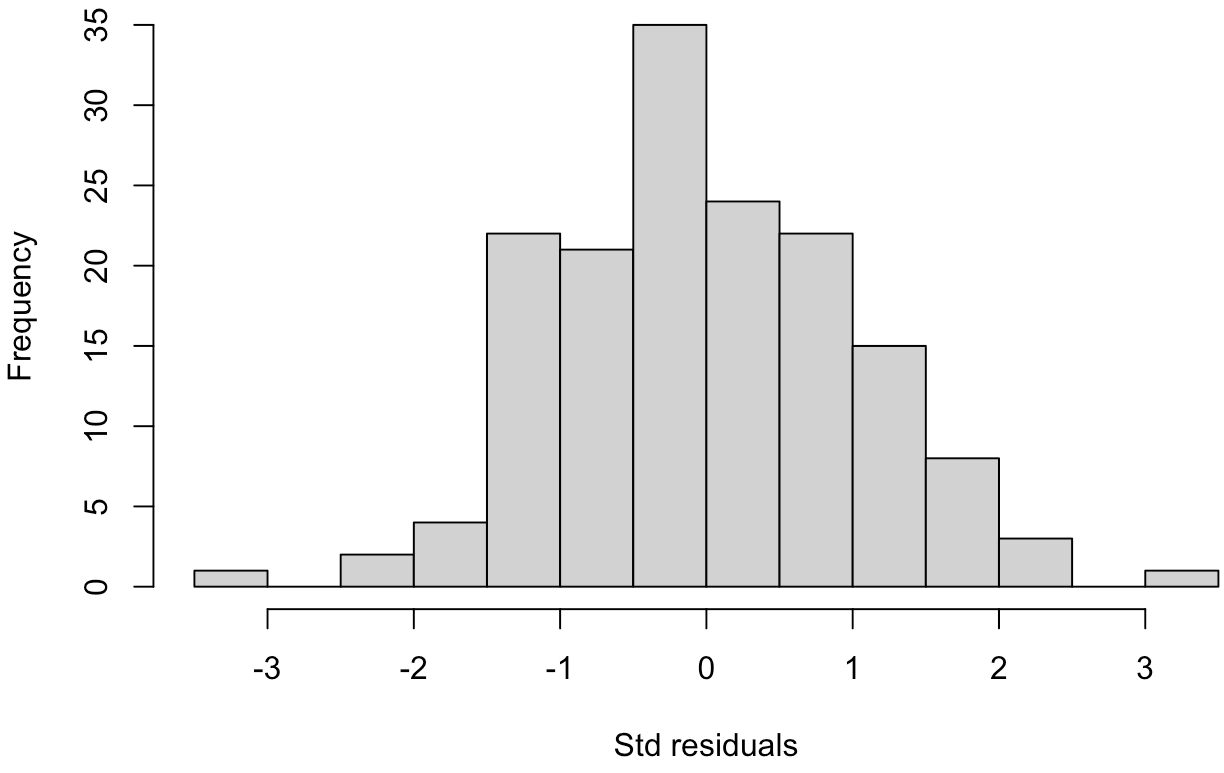
(d-3) Evaluate whether the deviance residuals are normally distributed by producing a QQ plot and histogram of the deviance residuals. What do you conclude?

```
model2.res <- resid(model2,type='deviance')

#QQplot
qqnorm(model2.res, ylab="Std residuals")
qqline(model2.res,col="blue",lwd=2)
```




```
#Histgram
hist(model2.res,10,xlab="Std residuals", main="")
```



Answer 4d-3:

The distribution of the QQplot improves as there is less skew on both tails. The histogram shows improvement as well as it shows more of a uniform distribution. Based on these plots the normality assumption should hold.

(d-4) Calculate the estimated dispersion parameter for this model. Is this an overdispersed model?

```
model2$deviance/model2$df.residual
```

```
## [1] 1.08233
```

Answer 4d-4:

The estimated dispersion parameter is now 1.08233 which is less than 2. The model is no longer overdispersed.

e. 4 pts - Overall, would you say model2 is a good-fitting model? If so, why? If not, what would you suggest to improve the fit and why? Note: We are not asking you to spend hours finding the best possible model but to offer plausible suggestions along with your reasoning.

Answer 4e:

Overall model2 is a good fitting model as shown by the results of deviance and pearson test. The model is no longer overdispersed and the normality assumption now holds. All of which is an improvement from model1. We can improve the fit of the model by removing variables that have no significance, such as Tenure.

Question 5: Prediction - 9 pts

Suppose there is an employee with the following characteristics:

1. **Age.Group**: 2
2. **Gender**: 0
3. **Tenure**: 2
4. **Num.Of.Products**: 2
5. **Is.Active.Member**: 1

a. 3 pts - Predict their probability of staying using model1.

Answer 5a:

```
preddata <- data.frame(Age.Group=2, Gender=0, Tenure=2, Num.Of.Products=2, Is.Active.Member=1)
preddata$Num.Of.Products<-as.factor(preddata$Num.Of.Products)
preddata$Age.Group<-as.factor(preddata$Age.Group)
preddata$Gender<-as.factor(preddata$Gender)
preddata$Is.Active.Member<-as.factor(preddata$Is.Active.Member)

pred.model1 <- predict(model1, preddata, type="response")
pred.model1
```

```
##         1
## 0.1997319
```

b. 3 pts - Predict their probability of staying using model2.

Answer 5b:

```
pred.model2 <- predict(model2, preddata, type="response")
pred.model2
```

```
##          1
## 0.08490958
```

c. 3 pts - Comment on how your predictions compare. i.e. which model is more reliable based on the analysis?

Answer 5c:

Model1 predictions are 0.1997319 and Model2 predictions are 0.08490958. Model2 is more reliable as it adds in other significant predictor variables in its predictions. Model2 is also a good fit based on our deviance analysis and the normality assumption holds true, whereas model1 does not.