ISYE 6414 Homework 3 Peer Assessment Solutions

ISYE 6414 Instructor

Summer 2021

Peer Grader Guidance

Please review the student expectations for peer review grading and peer review comments. Overall, we ask that you score with accuracy. When grading your peers, you will not only learn how to improve your future homework submissions but you will also gain deeper understanding of the concepts in the assignments. When assigning scores, consider the responses to the questions given your understanding of the problem and using the solutions as a guide. Moreover, please give partial credit for a concerted effort, but also be thorough. **Add comments to your review, particularly when deducting points, to explain why the student missed the points.** Ensure your comments are specific to questions and the student responses in the assignment.

Background

The owner of a company would like to be able to predict whether employees will stay with the company or leave. The data contains information about various characteristics of employees. See below for the description of these characteristics.

Data Description

The data consists of the following variables:

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

Note: Please do not treat any variables as categorical.

Read the data

```
# import the data
data = read.csv("hw4_data.csv", header=TRUE, fileEncoding="UTF-8-BOM")
data$Staying = data$Stay/data$Employees
head(data)
```

```
##   Age.Group Gender Tenure Num.Of.Products Is.Active.Member Stay Employees
## 1         2      1      3               1                0    5        11
## 2         2      1      4               1                0    5        10
## 3         2      1      4               1                1    2        13
## 4         2      0      7               1                0    3        10
## 5         2      1      7               1                0    2        14
```

```
## 6         2      0      4                2                 0    4        12
##     Staying
## 1 0.4545455
## 2 0.5000000
## 3 0.1538462
## 4 0.3000000
## 5 0.1428571
## 6 0.3333333
```

Question 1: Fitting a Model - 6 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. Call it **model1**.

(a) 2 pts - Display the summary of model1. What are the model parameters and estimates?

```
# Create model1
model1 = glm(Staying ~ Num.Of.Products, data=data, family='binomial', weights=Employees)
summary(model1)
```

```
##
## Call:
## glm(formula = Staying ~ Num.Of.Products, family = "binomial",
##     data = data, 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)       2.1457     0.1318   16.27   <2e-16 ***
## Num.Of.Products  -1.7668     0.1031  -17.13   <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
```

The model parameters are β_0 and β_1 and their estimates are:

$\hat{\beta_0} = 2.1457$

$\hat{\beta_1} = -1.7668$

Note that since there are no error terms, there is no parameter for the variance of the error (σ^2).

(b) 2 pts - Write down the equation for the odds of staying.

$$\frac{p_{staying}}{1 - p_{staying}} = e^{2.1457 - 1.7668 * X_{Num.Of.Products}}$$

(c) 2 pts - Provide a meaningful interpretation for the coefficient for *Num.Of.Products* with respect to the log-odds of staying and the odds of staying.

Log-odds:

- An one unit increase in *Num.Of.Products* decreases the log-odds of staying by 1.7668.

Odds:

- An one unit increase in *Num.Of.Products* decreases the odds of staying by 82.91%. $1 - e^{-1.7668} = 0.8291$
- An one unit increase in *Num.Of.Products* changes the odds of staying by a factor of 0.1709. $e^{-1.7668} = 0.1709$

Note: Only one correct interpretation with respect to the odds of staying is needed for credit.

Question 2: Inference - 9 pts

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

- Option 1:

```
# Calculate it by hand
cat("[",-1.7668-1.645*0.1031,",",-1.7668+1.645*0.1031,"]")
```

```
## [ -1.936399 , -1.597201 ]
```

- Option 2:

```
# Have R calculate the confidence interval using the Wald interval
confint.default(model1, "Num.Of.Products", level=0.9)
```

```
##                         5 %      95 %
## Num.Of.Products -1.936459 -1.597197
```

- Option 3:

```
# Have R calculate the confidence interval using profile-likelihood
confint(model1, "Num.Of.Products", level=0.9)
```

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

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

The 90% confidence interval for Num.Of.Products is ~ (-1.94,-1.60).

Note: Only one correct option is needed for credit. The above are different examples of possible answers.

(b) 3 pts - Is model1 significant overall? How do you come to your conclusion?

We will use a chi-square test to compare the fitted model to the null model.

```
# Calculate overall significance
1-pchisq((model1$null.dev - model1$deviance),
         (model1$df.null - model1$df.resid))
```

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

The p-value is very close to zero, indicating that the model is significant overall.

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

We look at the estimates and p-values given in the model summary.

We can see that the coefficients associated to the *intercept* and *Num.Of.Products* are significantly non-zero with p-values very close to zero.

To test for significantly negative coefficients, if the estimate is negative, and half the given p-value is below 0.01, then the p-value is significantly negative. The estimate for *Num.Of.Products* is negative, and the corresponding p-value is very close to zero. Half of the p-value of Num.Of.Products is still less than 0.01, indicating that Num.Of.Products is significantly negative.

Question 3: Goodness of fit - 9 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 residuals test
cat("Deviance residuals test p-value:",
1-pchisq(model1$deviance, model1$df.residual), end="\n")
```

```
## Deviance residuals test p-value: 0
```

```
# Pearson residuals test
pResid <- resid(model1, type = "pearson")
cat("Pearson residuals test p-value:",
1-pchisq(sum(pResid^2), model1$df.residual))
```

```
## Pearson residuals test p-value: 0
```

The p-values from both goodness of fit tests are close to 0, suggesting that we can reject the null hypothesis that the model is a good fit. These findings do not contradict those from Question 2b because it is possible for a model to have explanatory/predictive power while still being a poor fit.

(b) 3.5 pts - Perform visual analytics for checking goodness of fit for this model and write your observations. Be sure to address the model assumptions. Only deviance residuals are required for this question.

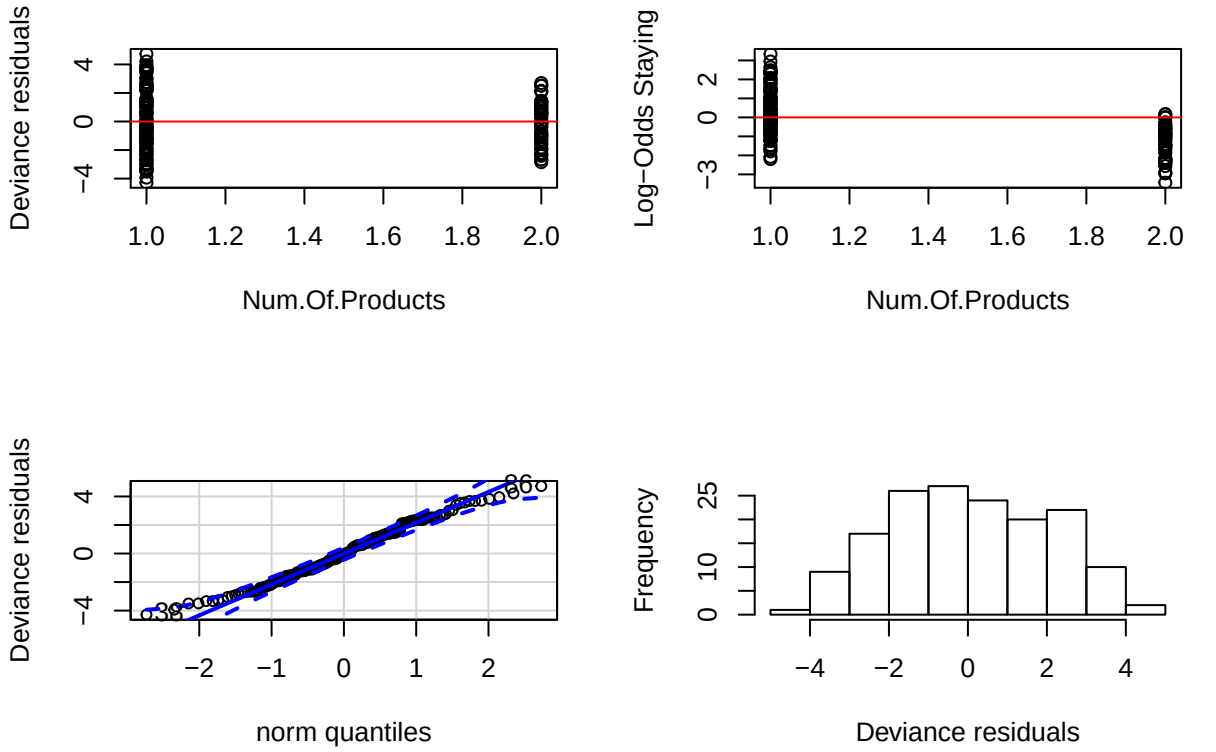
```
library(car)
```

```
## Loading required package: carData
```

```
par(mfrow=c(2,2))

# Store the deviance residuals
res = resid(model1,type="deviance")

# Plot the residuals against the predictor Num.of.Products
plot(data$Num.Of.Products,res, ylab="Deviance residuals", xlab="Num.Of.Products")
abline(0,0, col="red")

# Plot Logit vs. Num.of.Products
plot(data$Num.Of.Products, log(data$Staying/(1-data$Staying)),
ylab="Log-Odds Staying",xlab="Num.Of.Products")
abline(0,0, col="red")

# QQ-plot and Histogram
qqPlot(res, ylab="Deviance residuals")
```

```
## [1] 86 38
```

```
hist(res,10,xlab="Deviance residuals", main="")
```



Independence assumption: We can evaluate whether the residuals are uncorrelated or not by plotting the deviance residuals against our only predictor, *Num.Of.Products*. This plot does not show any clear pattern or clustering among the residuals.

Linearity assumption: We can assess the linearity assumption by plotting the log-odds of Staying against the predictor, Num.Of.Products. The predicting value, Num.Of.Products, only takes on two values, so evaluating linearity is not very practicable. There appears to be a general decreasing trend of the log-odds of Staying against Num.Of.Products.

When the model fits well, the deviance residuals should be approximately normal. We can assess this by using a QQ plot and a histogram of the deviance residuals. The qqplot shows that the majority of the deviance residuals fall inside of the 95% confidence interval. Based on these plots, model1 may be a good fit for the data.

Note: Interpretation of some of these plots can be subjective. Primarily, we're looking for sound justification and logical reasoning

(c) 2 pts - Calculate the dispersion parameter for this model. Is this an overdispersed model?

```
# Calculate overdispersion parameter
model1$deviance/model1$df.res
```

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

The estimated dispersion parameter is larger than 2, so there is overdispersion.

Question 4: Fitting the full model- 20 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. Call it **model2**.

```r
# Create model2
model2 = glm(Staying ~ Age.Group + Gender + Tenure + Num.Of.Products + Is.Active.Member,
             data=data,family='binomial', weights=Employees)
summary(model2)
```

```
##
## Call:
## glm(formula = Staying ~ Age.Group + Gender + Tenure + Num.Of.Products +
##     Is.Active.Member, family = "binomial", data = data, weights = Employees)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2638  -0.7662   0.0018   0.6836   2.8912
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.903330   0.330549  -5.758 8.51e-09 ***
## Age.Group         1.229014   0.075158  16.352  < 2e-16 ***
## Gender           -0.551438   0.093139  -5.921 3.21e-09 ***
## Tenure           -0.003574   0.016470  -0.217    0.828
## Num.Of.Products  -1.428767   0.111181 -12.851  < 2e-16 ***
## Is.Active.Member -0.871460   0.095034  -9.170  < 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: 171.94  on 152  degrees of freedom
## AIC: 604.66
##
## Number of Fisher Scoring iterations: 4
```

(a) 2.5 pts - Write down the equation for the probability of staying.

$$p_{staying} = \frac{e^{-1.903+1.229*x_{Age.Group}-0.551*x_{Gender}-0.004*x_{Tenure}-1.429*x_{Num.Of.Products}-0.871*x_{Is.Active.Member}}}{1+e^{-1.903+1.229*x_{Age.Group}-0.551*x_{Gender}-0.004*x_{Tenure}-1.429*x_{Num.Of.Products}-0.871*x_{Is.Active.Member}}}$$

(b) 2.5 pts - Provide a meaningful interpretation for the coefficients of *Age.Group* and *Is.Active.Member* with respect to the odds of staying.

Age.Group:

- A one unit increase in *Age.Group* increases the odds of staying by 242% holding all other predictors constant. $e^{1.229} - 1 = 2.418$
- A one unit increase in *Age.Group* changes the odds of staying by a factor of 3.418 holding all other predictors constant. $e^{1.229} = 3.418$

Is.Active.Member:

- A one unit increase in *Is.Active.Member* decreases the odds of staying by 58.1% holding all other predictors constant. $1 - e^{-0.871} = 0.581$
- A one unit increase in *Is.Active.Member* changes the odds of staying by a factor of 0.419 holding all other predictors constant. $e^{-0.871} = 0.419$

(c) 2.5 pts - Is *Is.Active.Member* significant given the other variables in model2?

Is.Active.Member has a p-value less than 2e-16. Since this p-value is so small, Is.Active.Member is significant given the other variables in the model.

(d) 10 pts - Has your goodness of fit been affected? Repeat the tests, plots, and dispersion parameter calculation you performed in Question 3 with model2.

First, let's check algorithmically whether the model seems to be a good fit by performing a goodness of fit test using Pearson and Deviance residuals.

```
# Deviance residuals test
cat("Deviance residuals test p-value:",
1-pchisq(model2$deviance, model2$df.residual), end="\n")
```

```
## Deviance residuals test p-value: 0.1282109
```

```
# Pearson residuals test
pResid <- resid(model2, type = "pearson")
cat("Pearson residuals test p-value:",
1-pchisq(sum(pResid^2), model2$df.residual))
```

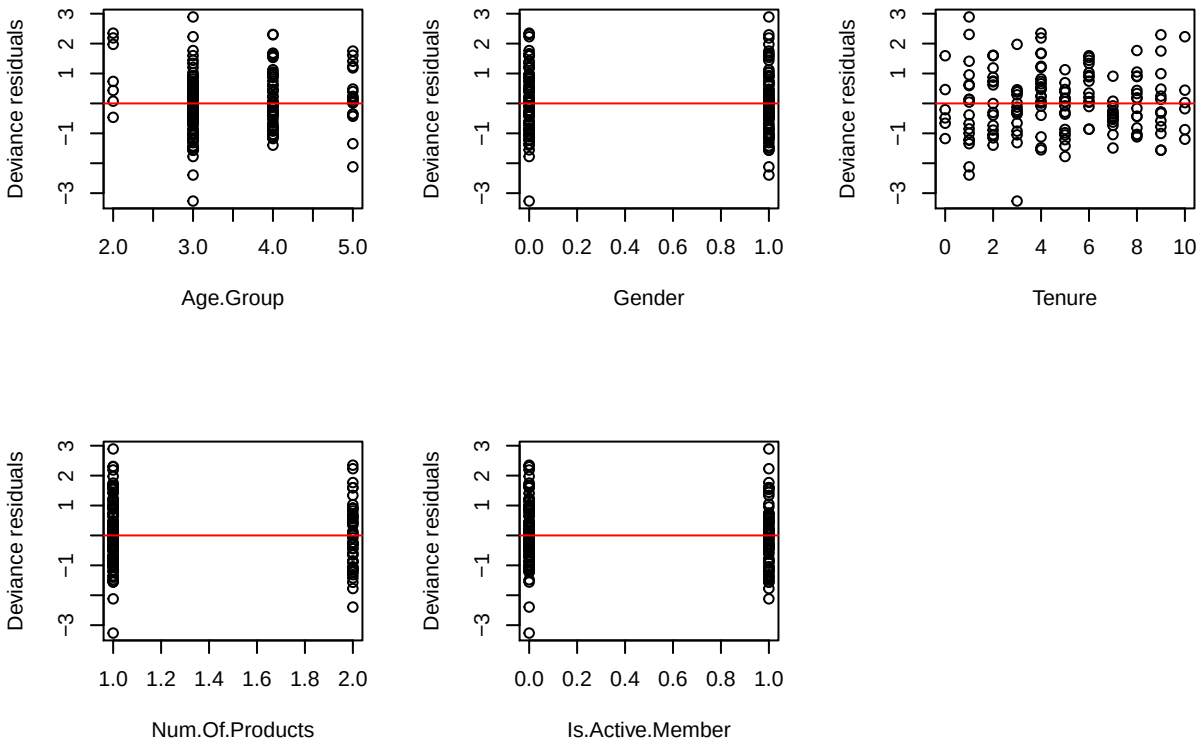
```
## Pearson residuals test p-value: 0.200838
```

The p-values of both tests for goodness of fit are large, suggesting that the model might be a good fit.

Next, let's use visual analytics for assessing the model assumptions.

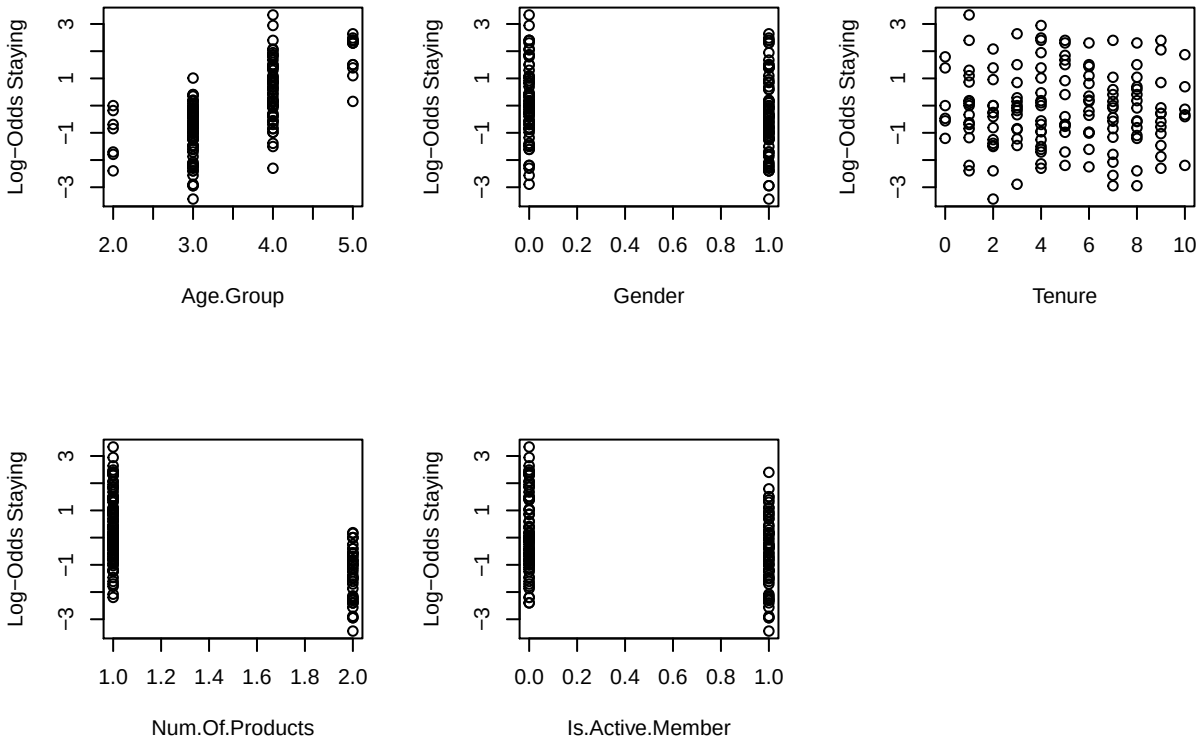
Independence assumption: We can evaluate whether the residuals are uncorrelated or not by plotting the deviance residuals against the five predictors. We could also use the plot of residuals vs. fitted values. Overall, none of the plots show any clear pattern or clustering among the residuals.

```
# Store the deviance residuals
res2 = resid(model2,type="deviance")
# Grid the plots
par(mfrow=c(2,3))
# Plot the residuals against the predictor Age.Group
plot(data$Age.Group,res2,ylab="Deviance residuals",xlab="Age.Group")
abline(0,0, col="red")
# Plot the residuals against the predictor Gender
plot(data$Gender,res2,ylab="Deviance residuals",xlab="Gender")
abline(0,0, col="red")
# Plot the residuals against the predictor Tenure
plot(data$Tenure,res2,ylab="Deviance residuals",xlab="Tenure")
abline(0,0, col="red")
# Plot the residuals against the predictor Num.Of.Products
plot(data$Num.Of.Products,res2,ylab="Deviance residuals",xlab="Num.Of.Products")
abline(0,0, col="red")
# Plot the residuals against the predictor Is.Active.Member
plot(data$Is.Active.Member,res2,ylab="Deviance residuals",xlab="Is.Active.Member")
abline(0,0, col="red")
```



Linearity assumption: We can assess the linearity assumption by plotting the log-odds of Staying against the predictors. Most of the predictors take on few values, so evaluating linearity is not very practicable. There appears to be a general increasing trend of the log-odds of Staying against Age.Group, as well as a general decreasing trend of the log-odds of Staying against Num.Of.Products. There does not appear to be a strong linear relationship between the log-odds of Staying and Tenure.

```
# Grid the plots
par(mfrow=c(2,3))
# Plot the Log-Odds Staying against the predictor Age.Group
plot(data$Age.Group, log(data$Staying/(1-data$Staying)),ylab="Log-Odds Staying",
     xlab="Age.Group")
# Plot the Log-Odds Staying against the predictor Gender
plot(data$Gender, log(data$Staying/(1-data$Staying)),
ylab="Log-Odds Staying", xlab="Gender")
# Plot the Log-Odds Staying against the predictor Tenure
plot(data$Tenure, log(data$Staying/(1-data$Staying)),
ylab="Log-Odds Staying", xlab="Tenure")
# Plot the Log-Odds Staying against the predictor Num.Of.Products
plot(data$Num.Of.Products, log(data$Staying/(1-data$Staying)),
ylab="Log-Odds Staying", xlab="Num.Of.Products")
# Plot the Log-Odds Staying against the predictor Is.Active.Member
plot(data$Is.Active.Member, log(data$Staying/(1-data$Staying)),
ylab="Log-Odds Staying", xlab="Is.Active.Member")
```

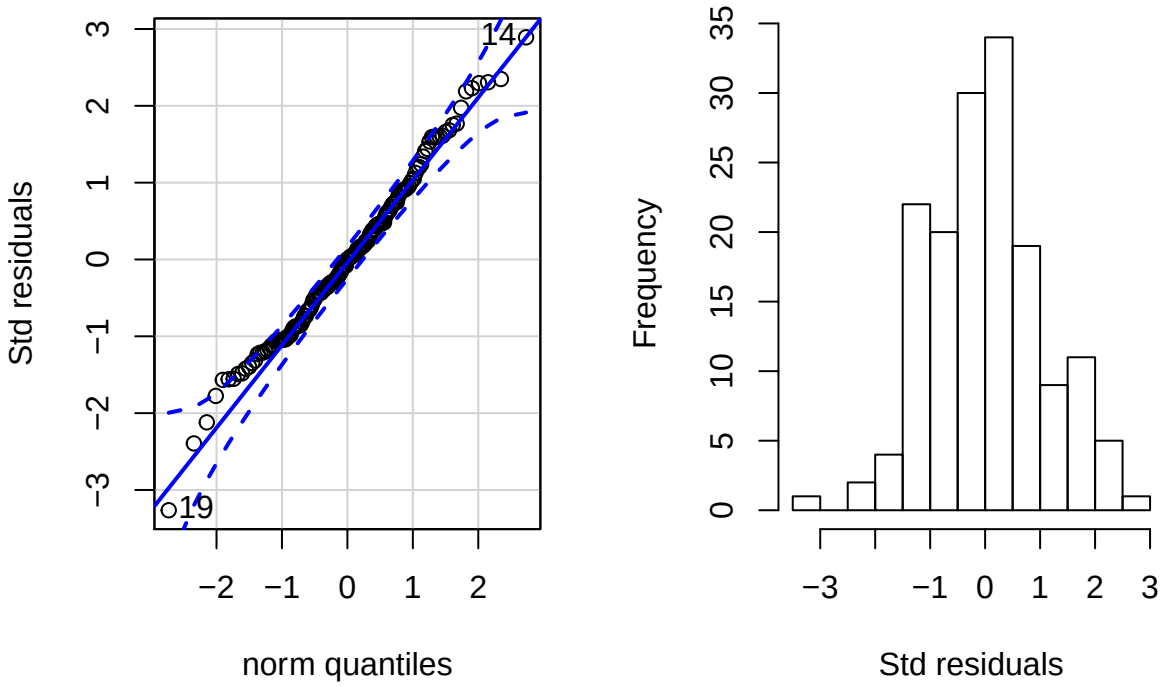


Additionally, we can evalaute the QQ plot and the histogram of the deviance residuals. When the model fits well, the deviance residuals should be approximately normal. The qqplot shows that the majority of the deviance residuals fall inside of the 95% confidence interval. Based on these plots, model2 seems to be a good fit for the data.

```
# QQ-plot and Histogram
par(mfrow=c(1,2))
qqPlot(res2, ylab="Std residuals")
```

```
## [1] 19 14
```

```
hist(res2,10,xlab="Std residuals", main="")
```



Finally, let's calculate the dispersion parameter.

```
model2$deviance/model2$df.res
```

```
## [1] 1.131172
```

The dispersion parameter is now smaller than 2. There is no overdispersion anymore.

Note: Interpretation of some of these plots can be subjective. Primarily, we're looking for sound justification and logical reasoning

(e) 2.5 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.

From the plots and tests in the previous question, the goodness of fit appears to have improved. Overall, model2 appears to be a good-fitting model.

One possibility to improve the fit is to change the link function. Changing it from logit to cloglog (complementary log-log) reduces the deviance slightly (from 171.94 to 167.21).

Some other ideas for improving the fit are transforming the data, removing influential points from the data, and considering other predicting variables.

```
# Create model3
model3 = glm(Staying ~ Age.Group + Gender + Tenure + Num.Of.Products + Is.Active.Member,
data=data,family=binomial(link="cloglog"), weights=Employees)
dp <-sum(residuals(model3,type="deviance")^2)/model3$df.res
summary(model3, dispersion=dp)
```

```
##
## Call:
## glm(formula = Staying ~ Age.Group + Gender + Tenure + Num.Of.Products +
##     Is.Active.Member, family = binomial(link = "cloglog"), data = data,
##     weights = Employees)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -3.1335  -0.6421  -0.1019   0.6338   2.6734
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.456824   0.242084  -6.018 1.77e-09 ***
## Age.Group         0.796179   0.049123  16.208  < 2e-16 ***
## Gender           -0.360868   0.065145  -5.539 3.03e-08 ***
## Tenure           -0.004094   0.011478  -0.357    0.721
## Num.Of.Products  -1.092100   0.094629 -11.541  < 2e-16 ***
## Is.Active.Member -0.612588   0.068478  -8.946  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1.100078)
##
##     Null deviance: 981.04  on 157  degrees of freedom
## Residual deviance: 167.21  on 152  degrees of freedom
## AIC: 599.93
##
## Number of Fisher Scoring iterations: 4
```

Question 5: Prediction - 6 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) 2 pts - Predict their probability of staying using model1.

```
# Create the new employee
newemployee = data.frame(Age.Group=2, Gender=0, Tenure=2,
                         Num.Of.Products=2, Is.Active.Member=1)
predict(model1, newemployee, type="response")
```

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

Using model1, given the above characteristics of an employee the probability of staying is predicted to be equal to 0.1997319 (~20%).

(b) Predict their probability of staying using model2.

```
predict(model2,newemployee,type="response")
```

```
##          1
## 0.03987005
```

Using model2, given the above characteristics of an employee the probability of staying is predicted to be equal to 0.03987005(~4%).

(c) Comment on how your predictions compare.

When Age.Group, Gender, Tenure, and Is.Active.Member are taken into consideration, the employee's predicted probability of staying at the company decreases by about 0.16. Based on the goodness of fit tests, model2 seems to be much more reliable than model1. However, we might need to split our data set into training and testing sets and calculate prediction accuracy measurements in order to further evaluate the prediction accuracy of the models.