HW3 Peer Assessment

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("hw3_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
```

```
attach(data)
```

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?

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

```
##
## Call:
## glm(formula = Staying ~ Num.Of.Products, family = binomial, 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
```

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

$\frac{p_{Staying}}{(1-p_{Staying})} = e^{(2.1457-1.7668*Num.Of.Products)}$

(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.

For each unit increase in *Num.Of.Products*

1. Log Odds of staying decreases by 1.7668.
2. Odds of staying decreases by 5.85 ($e^{1.7668}$)

Question 2: Inference - 9 pts

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

```
#exp(confint(model1, level = 0.90))
# -1.7668
confint(model1, level = 0.90)
```

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

```
##                        5 %      95 %
## (Intercept)      1.930071  2.363889
## Num.Of.Products -1.938361 -1.598965
```

```
## Lower summary(model1)$coefficients[2,1] - (summary(model1)$coefficients[2,2] * qnorm(0.95))
## Upper summary(model1)$coefficients[2,1] + (summary(model1)$coefficients[2,2] * qnorm(0.95))
```

Answer : 90% CI for Num.Of.Products = -1.938361 -1.598965

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

Answer : Without the residual analysis, it is difficult to say if model1 is significant overall. But the coefficients are statistically significant

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

both the coefficient, Intercept and Num.Of.Products are significant at significance level 0.01. 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.

```
## Test for overall regression
gstat = model1$null.deviance - deviance(model1)
c(gstat, 1-pchisq(gstat,length(coef(model1))-1))
```

```
## [1] 348.996   0.000
```

```
## Test for GOF: Using deviance residuals
c(deviance(model1), 1 - pchisq(deviance(model1),156))
```

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

```
## Test for GOF: Using Person residuals
pearres1 = residuals(model1,type="pearson")
pearson1.tvalue = sum(pearres1^2)
c(pearson1.tvalue, 1-pchisq(pearson1.tvalue,156))
```

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

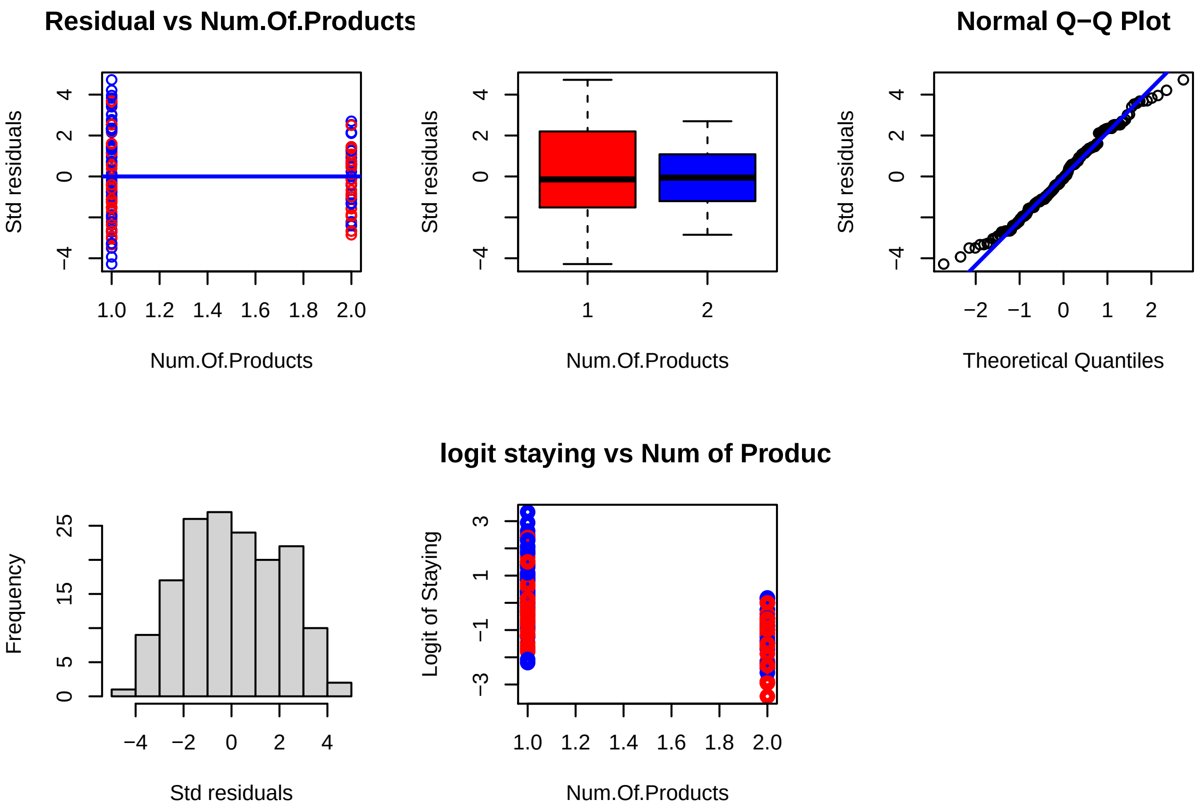
Answer : Using deviance residuals: P-value = 0.00. Reject the null hypothesis of good fit (thus NOT a good fit) Using Pearson residual: P value = 0.0000. Reject the null hypothesis of good fit (thus NOT a good fit)

Comparing to Answer 2b, the model might not satisfy all the assumptions

(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.

```
par(mfrow=c(2,3))
#Residual Analysis
res1 = resid(model1,type="deviance")
plot(Num.Of.Products,res1,ylab="Std residuals",xlab="Num.Of.Products",
     main="Residual vs Num.Of.Products", col=c("red","blue"))
abline(0,0,col="blue",lwd=2)
boxplot(res1~Num.Of.Products,ylab = "Std residuals", col=c("red","blue"))
#Normality Assumption
qqnorm(res1, ylab="Std residuals")
qqline(res1,col="blue",lwd=2)
hist(res1,10,xlab="Std residuals", main="")

#Linearity Assumption
plot(Num.Of.Products,log((Staying)/(1-Staying)), ylab="Logit of Staying",
     main="logit staying vs Num of Products", col=c("red","blue"),lwd=3)
```



Answer : Based on the residual analysis.. looks like the Normality assumptions holds though the distributions has long tails. not able to determine linearity assumption wrt to *Num.Of.Products* as there are only 2 values for the predictor. But the variance for each of the 2 values is not consistent. Higher variance in data for 1 compared to 2

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

```
d = deviance(model1)
d2 = sum(residuals(model1, type = "deviance")^2)
n = nrow(data)
p = length(model1$coefficients) - 1

disp = d2/(n - p - 1)
disp
```

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

Answer : the Overdispersion $\hat{\phi}$= 4.05.. this is greater than 2.. so looks like 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**.

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

```
##
## Call:
```

```
## glm(formula = Staying ~ Age.Group + Gender + Tenure + Num.Of.Products +
##     Is.Active.Member, family = binomial, 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.

Answer :

$$p_{Staying} = \frac{e^{(-1.903330+1.229014*Age.Group-0.551438*Gender-0.003574*Tenure-1.428767*Num.Of.Products-0.871460*Is.Active.Member)}}{(1+e^{(-1.903330+1.229014*Age.Group-0.551438*Gender-0.003574*Tenure-1.428767*Num.Of.Products-0.871460*Is.Active.Member)})}$$

(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.

Answer :

Odds of staying *increases* by a factor of 3.418 ($e^{1.229014}$) for one unit increase in *Age.Group*

Odds of staying *decreases* by a factor of 2.39 ($\frac{1}{e^{-0.871460}}$) for one unit increase in *Is.Active.Member*.

The above assumptions are based on if all other predictors are constant

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

Answer :

the p-value for *Is.Active.Member* is < 2e-16. this means the *Is.Active.Member* is statistically significant

(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.

```
c(deviance(model2), 1 - pchisq(deviance(model2),152))
```

```
## [1] 171.9381966   0.1282109
```

```
pearres2 = residuals(model2,type="pearson")
pearson2.tvalue = sum(pearres2^2)
c(pearson2.tvalue, 1-pchisq(pearson2.tvalue,152))
```

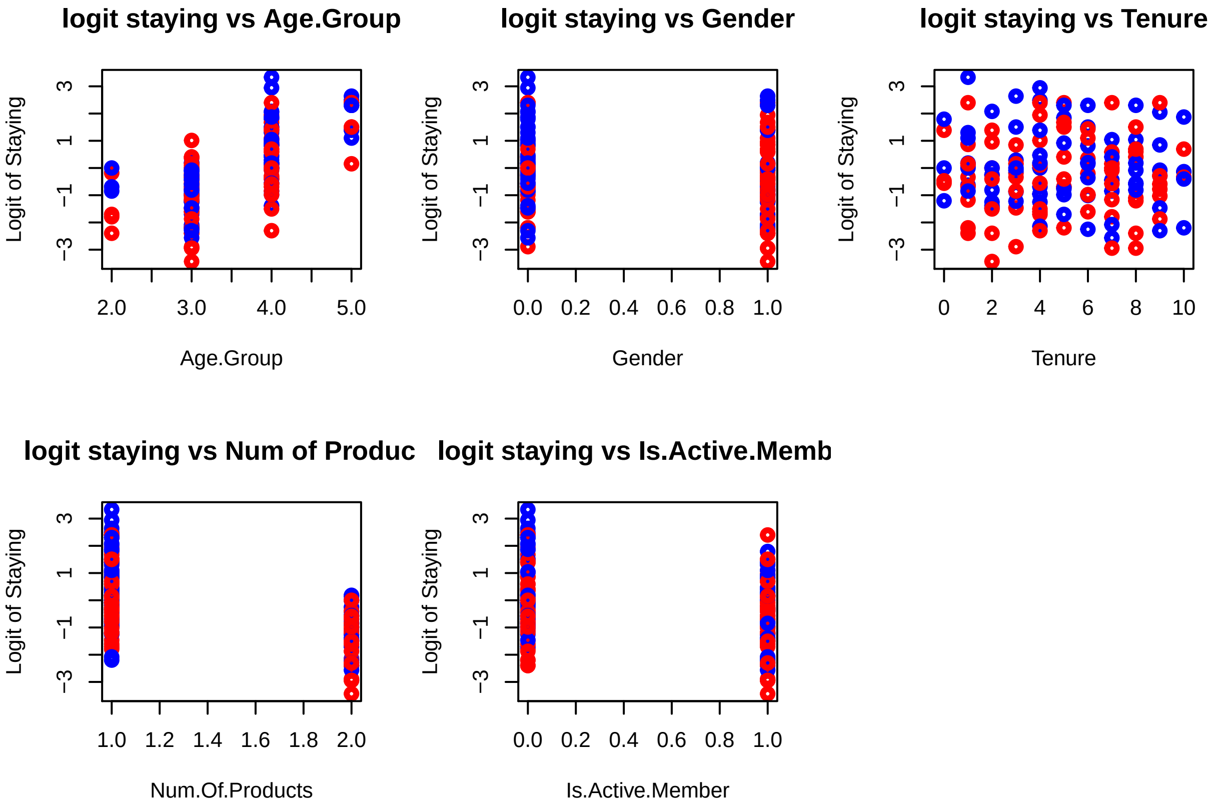
```
## [1] 166.390888   0.200838
```

Answer :

1. Using deviance residuals : p-value = 0.1282109
2. Using Pearson residuals : p-value = 0.200838

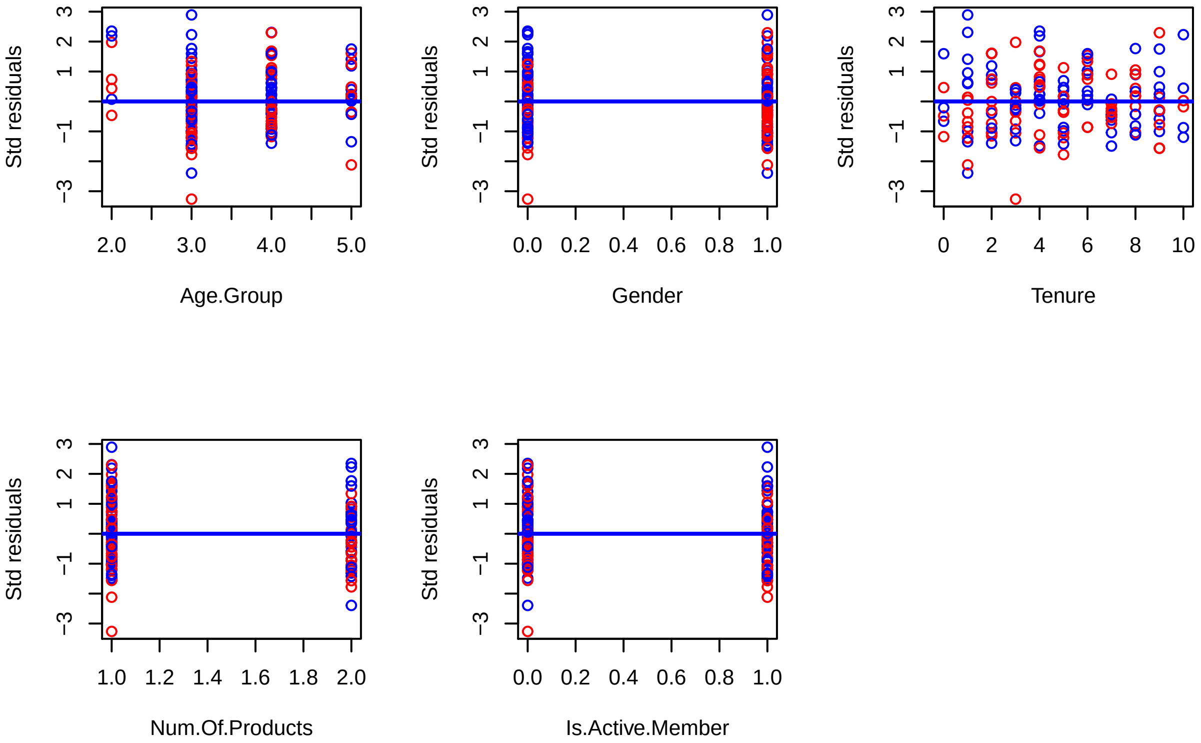
Thus we cannot reject the null hypothesis of good fit using either Pearson residuals or Deviance residuals. Based on this the models seems a good fit.

```
#Linearity Assumptions
par(mfrow=c(2,3))
plot(Age.Group,log((Staying)/(1-Staying)), ylab="Logit of Staying", xlab = "Age.Group",
     main="logit staying vs Age.Group", col=c("red","blue"),lwd=3)
plot(Gender,log((Staying)/(1-Staying)), ylab="Logit of Staying",xlab = "Gender",
     main="logit staying vs Gender", col=c("red","blue"),lwd=3)
plot(Tenure,log((Staying)/(1-Staying)), ylab="Logit of Staying",xlab = "Tenure",
     main="logit staying vs Tenure", col=c("red","blue"),lwd=3)
plot(Num.Of.Products,log((Staying)/(1-Staying)), ylab="Logit of Staying",xlab = "Num.Of.Products",
     main="logit staying vs Num of Products", col=c("red","blue"),lwd=3)
plot(Is.Active.Member,log((Staying)/(1-Staying)), ylab="Logit of Staying",xlab = "Is.Active.Member",
     main="logit staying vs Is.Active.Member", col=c("red","blue"),lwd=3)
```

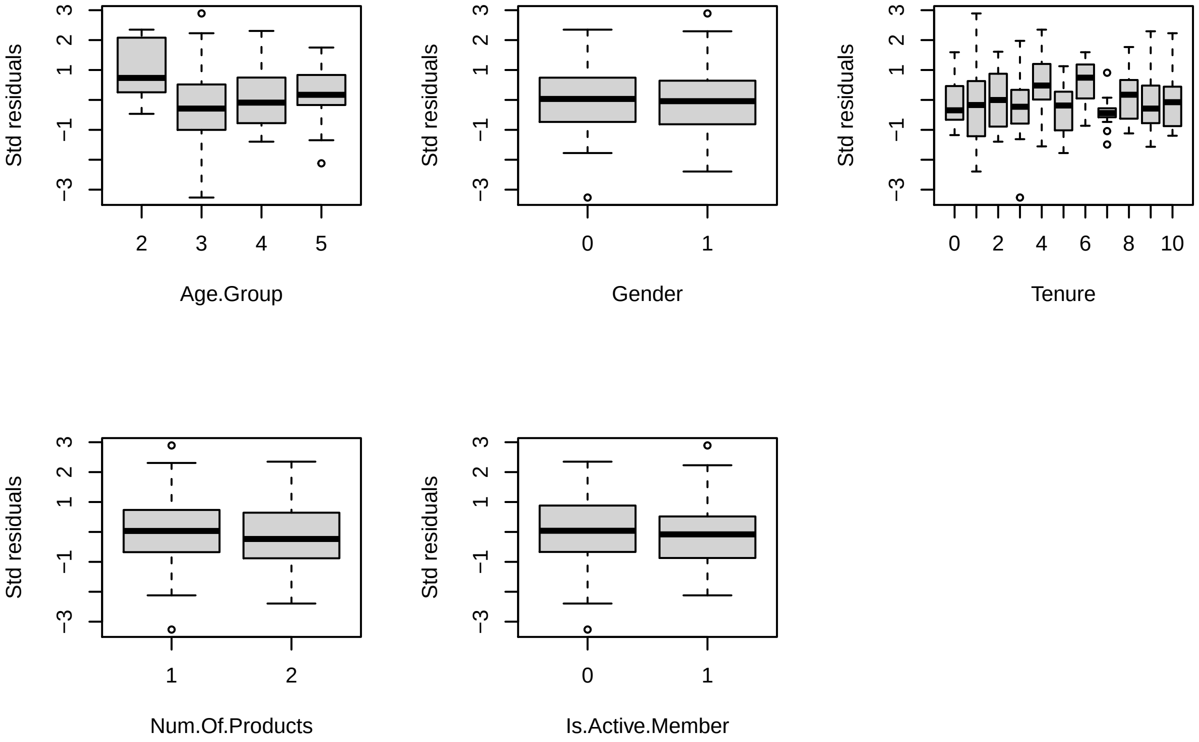


```
#Residual Analysis
par(mfrow=c(2,3))
res2 = resid(model2,type="deviance")
plot(Age.Group,res2,ylab="Std residuals",xlab="Age.Group", col=c("red","blue"))
abline(0,0,col="blue",lwd=2)
plot(Gender,res2,ylab="Std residuals",xlab="Gender", col=c("red","blue"))
abline(0,0,col="blue",lwd=2)
plot(Tenure,res2,ylab="Std residuals",xlab="Tenure", col=c("red","blue"))
abline(0,0,col="blue",lwd=2)
```

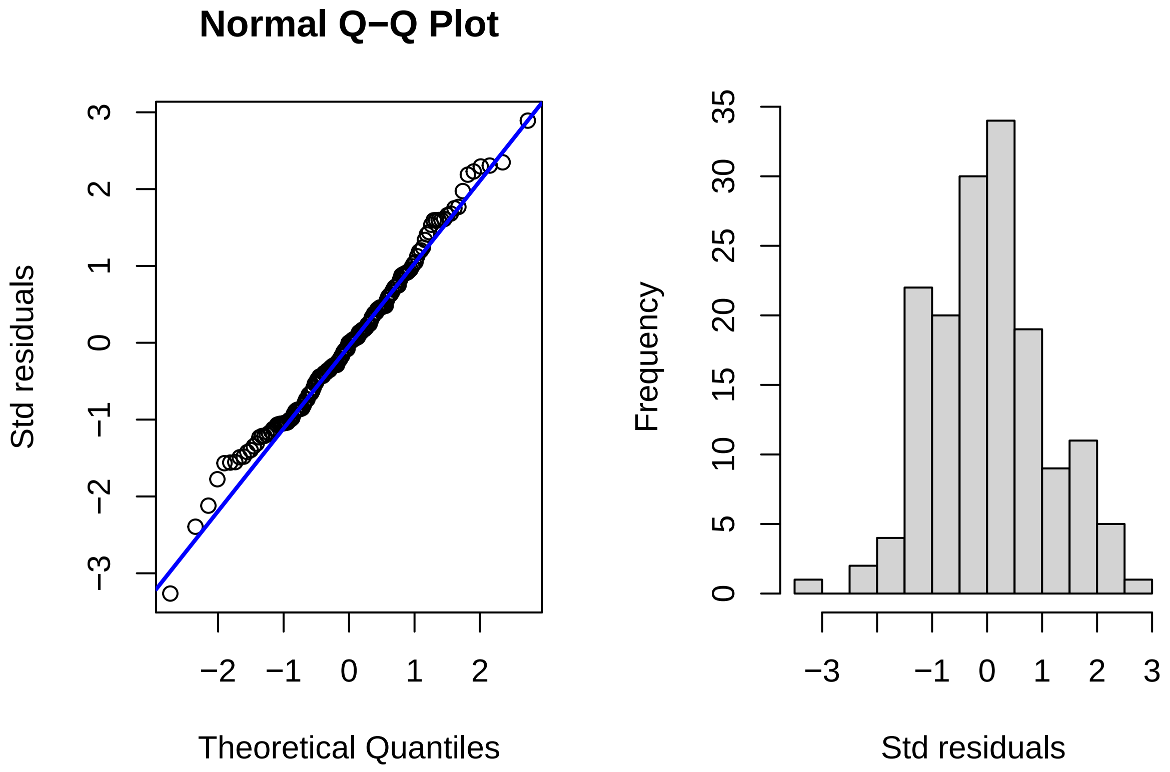
```
plot(Num.Of.Products,res2,ylab="Std residuals",xlab="Num.Of.Products", col=c("red","blue"))
abline(0,0,col="blue",lwd=2)
plot(Is.Active.Member,res2,ylab="Std residuals",xlab="Is.Active.Member", col=c("red","blue"))
abline(0,0,col="blue",lwd=2)
```



```
par(mfrow=c(2,3))
boxplot(res2~Age.Group,ylab = "Std residuals", xlab = "Age.Group")
boxplot(res2~Gender,ylab = "Std residuals", xlab = "Gender")
boxplot(res2~Tenure,ylab = "Std residuals", xlab = "Tenure")
boxplot(res2~Num.Of.Products,ylab = "Std residuals", xlab = "Num.Of.Products")
boxplot(res2~Is.Active.Member,ylab = "Std residuals", xlab = "Is.Active.Member")
```



```
#Normality Assumption
par(mfrow=c(1,2))
qqnorm(res2, ylab="Std residuals")
qqline(res2,col="blue",lwd=2)
hist(res2,10,xlab="Std residuals", main="")
```



```
#Calculate dispersion
d = deviance(model2)
d2 = sum(residuals(model2, type = "deviance")^2)
n = nrow(data)
p2 = length(model2$coefficients) - 1

disp = d2/(n - p2 - 1)
disp
```

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

(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.

Answer :

1. Using deviance residuals : p-value = 0.1282109
2. Using Pearson residuals : p-value = 0.200838 Thus we cannot reject the null hypothesis of good fit using either Pearson residuals or Deviance residuals. Based on this the models seems a good fit.
3. the dispersion $\hat{\phi} = 1.131..$ This is less than 2 .. so the model does not have overdispersion.
4. Tenure and Age.Group have different means across values whereas Age.Group and Is.Active.Member have very similar means

We can try and improve the fit by treating the variables as factors.

```
model3 = glm(Staying ~ as.factor(Age.Group)+as.factor(Gender)+as.factor(Tenure)+
               as.factor(Num.Of.Products)+as.factor(Is.Active.Member),
             weights=Employees , family=binomial)
summary(model3)
```

```
##
## Call:
## glm(formula = Staying ~ as.factor(Age.Group) + as.factor(Gender) +
##     as.factor(Tenure) + as.factor(Num.Of.Products) + as.factor(Is.Active.Member),
##     family = binomial, weights = Employees)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.78712  -0.71684  -0.01341   0.65484   3.15512
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -0.0904612  0.3727914  -0.243    0.808
## as.factor(Age.Group)3        0.3261427  0.2760093   1.182    0.237
## as.factor(Age.Group)4        1.6809205  0.2774345   6.059 1.37e-09 ***
## as.factor(Age.Group)5        2.8837910  0.3442932   8.376  < 2e-16 ***
## as.factor(Gender)1          -0.5749022  0.0944812  -6.085 1.17e-09 ***
## as.factor(Tenure)1           0.0021078  0.2835698   0.007    0.994
## as.factor(Tenure)2           0.0994498  0.2837427   0.350    0.726
## as.factor(Tenure)3          -0.1366218  0.2863104  -0.477    0.633
## as.factor(Tenure)4           0.2213202  0.2868505   0.772    0.440
## as.factor(Tenure)5          -0.0538657  0.2901980  -0.186    0.853
## as.factor(Tenure)6           0.3646101  0.2869467   1.271    0.204
## as.factor(Tenure)7          -0.2742701  0.2892822  -0.948    0.343
## as.factor(Tenure)8             0.0292250  0.2924315   0.100    0.920
## as.factor(Tenure)9            -0.0006682  0.2859176  -0.002    0.998
## as.factor(Tenure)10            0.0428148  0.3629838   0.118    0.906
## as.factor(Num.Of.Products)2  -1.4465708  0.1140149 -12.688  < 2e-16 ***
## as.factor(Is.Active.Member)1 -0.8546692  0.0964427  -8.862  < 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: 149.10  on 141  degrees of freedom
## AIC: 603.83
##
## Number of Fisher Scoring iterations: 4
```

```
c(deviance(model3), 1 - pchisq(deviance(model3),141))
```

```
## [1] 149.1045605   0.3039067
```

```
pearres3 = residuals(model3,type="pearson")
pearson3.tvalue = sum(pearres3^2)
c(pearson3.tvalue, 1-pchisq(pearson3.tvalue,141))
```

```
## [1] 141.3209749   0.4765538
```

```
d = deviance(model3)
d3 = sum(residuals(model3, type = "deviance")^2)
n = nrow(data)
p3 = length(model3$coefficients) - 1

disp = d3/(n - p3 - 1)
disp
```

```
## [1] 1.057479
```

Answer : changing all the predictors as factors give a higher p-value compared to the model2. looks like the model also holds the assumptions better. Over dispersion is close to 1

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.

```
newData1 = data.frame(Num.Of.Products=2)
predict.glm(model1, newData1)
```

```
##         1
## -1.387971
```

Answer :

$p_{staying} = \frac{e^{-1.387971}}{1+e^{-1.387971}}$ $p_{staying} = 0.199$

(b) 2 pts - Predict their probability of staying using model2.

```
newData2 = data.frame(Age.Group=2, Gender=0, Tenure=2, Num.Of.Products=2, Is.Active.Member=1)
predict.glm(model2, newData2)
```

```
##         1
## -3.181443
```

Answer :

$p_{staying} = \frac{e^{-3.181443}}{1+e^{-3.181443}}$ $p_{staying} = 0.04$

(c) 2 pts - Comment on how your predictions compare.

Model 1 the probably of staying is higher compared to the model2