HW4 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

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.Gro…	Gen…	Ten…	Num.Of.Products	Is.Active.Member	S…	Employe…	Staying
	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<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

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? Intercept and Num.of.Products are the two parameters here and their estimates are 2.1457 and -1.7668 respectively.

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

(b) 2 pts - Write down the equation for the odds of staying.
odds of staying = e^(2.1457-1.7668) = 1.46

(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.

The log-odds of staying decrees by 1.7668 with 1 unit increase of “Num.of.Products”.
The odds of staying decreases by 82.91% (1-e^(-1.7668)) = 1-0.1708 = 0.8291) with 1 unit increase of “Num.of.Products”.
or The odds of staying changes by a factor of 0.17 (e^(-1.7668) = 0.17) with 1 unit increase of “Num.of.Products”.

Question 2: Inference - 9 pts

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

```
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
```

(b) 3 pts - Is model1 significant overall? How do you come to your conclusion?
The p-value of below test is “0” very low , thus the model1 is statistically significant overall.

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

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

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

The “Num.Of.Products” and “intercept” both are nonzero at the 0.01 significance level. The Num.of.Products is the significantly negative coefficient as the p-value(2e-16) is approximate 0.

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.

Using both deviance and Pearson residuals we can reject the null hypothesis of good fit due to p-value being approximately “0”. Hence we can say that the model is not a good fit.

From the question 2(b) we concluded that the overall model is statistically significant but, from the deviance and Pearson residuals we concluded that the model is not a good fit.

```
c(deviance(model1), 1-pchisq(deviance(model1),156))
```

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

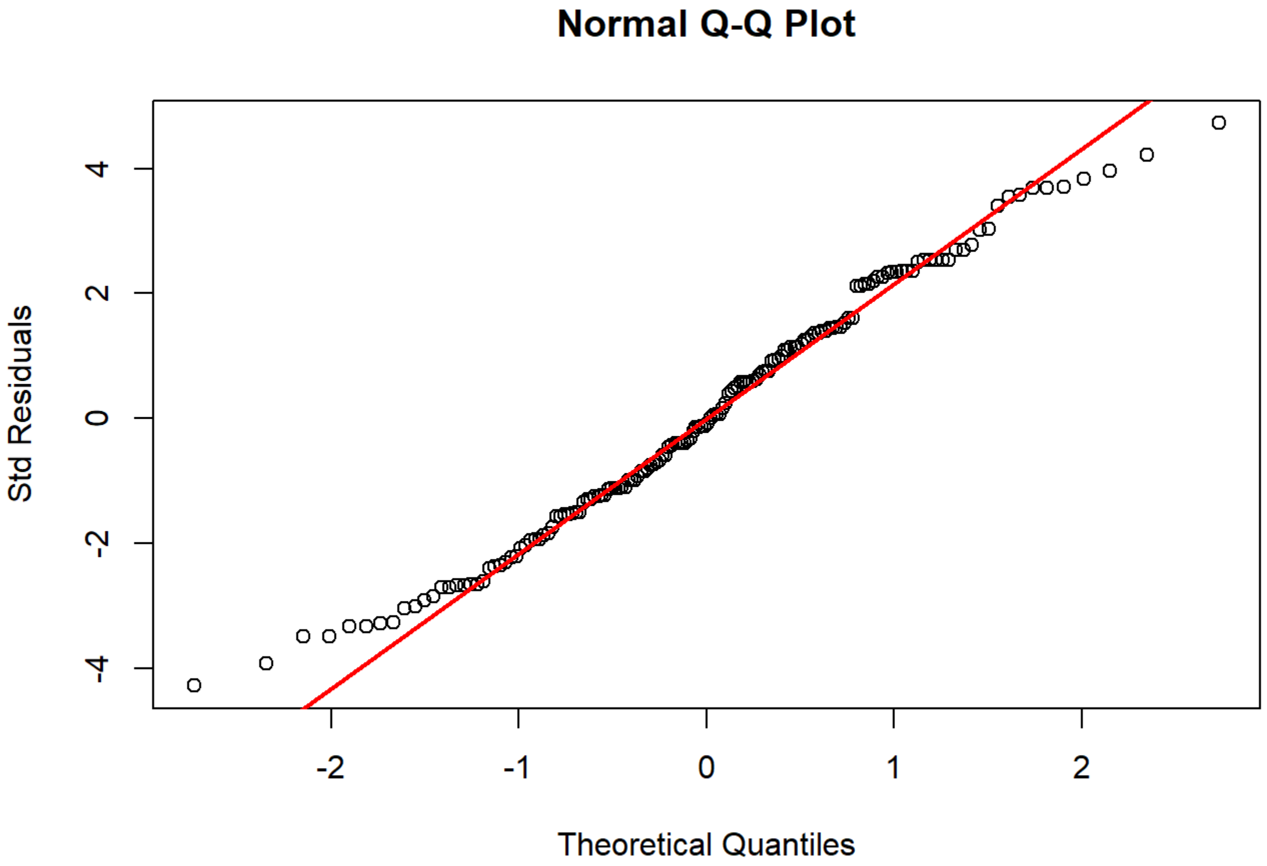
```
pearson1 <- residuals(model1, type = "pearson")
pearson.tvalue <- sum(pearson1^2)
c(pearson.tvalue, 1-pchisq(pearson.tvalue,156))
```

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

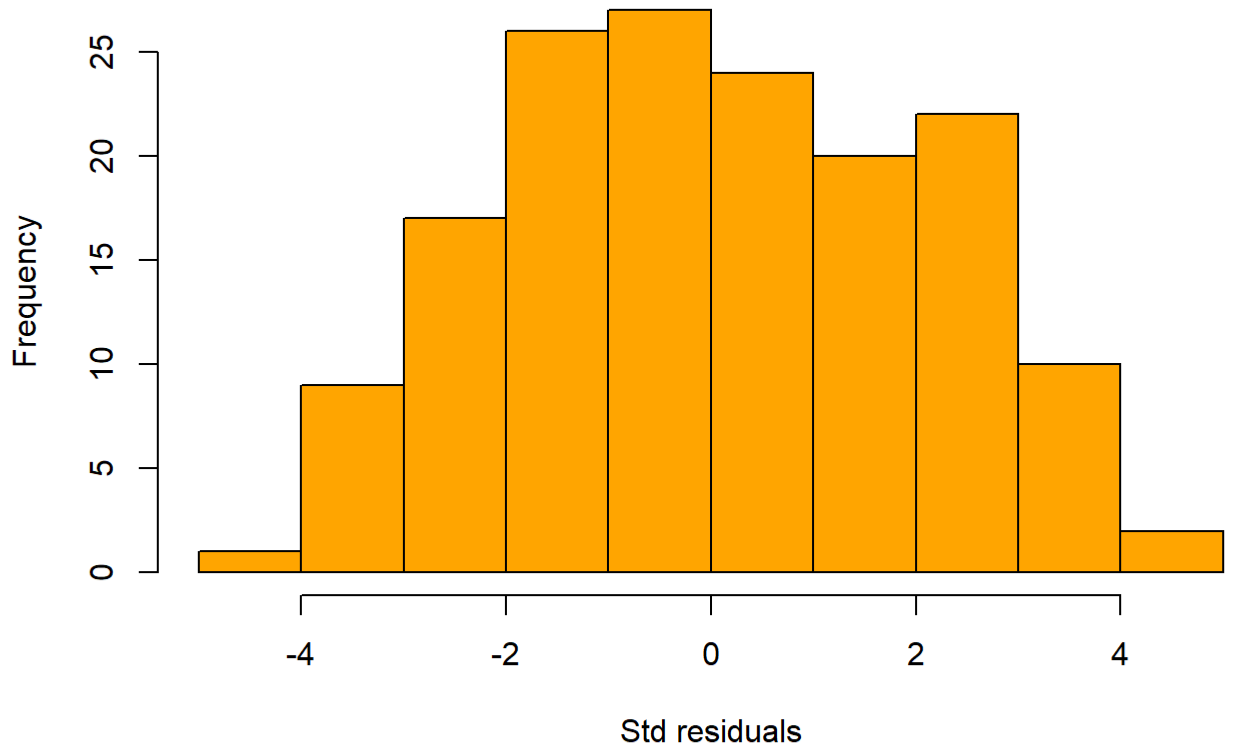
(b) 3.5 pts - Perform visual analytics for checking goodness of fit for this model and write your observations. Be sure to address the model assumptions. Only deviance residuals are required for this question.

The QQ plot shows slight heavy tails and histogram is also slightly skewed. Hence the model is not good fit. Overall we can say that normality assumption does not hold.

```
res <- resid(model1, type = "deviance")
#par(mfrow = c(2,1))
qqnorm(res, ylab = "Std Residuals")
qqline(res,col="red",lwd =2)
```

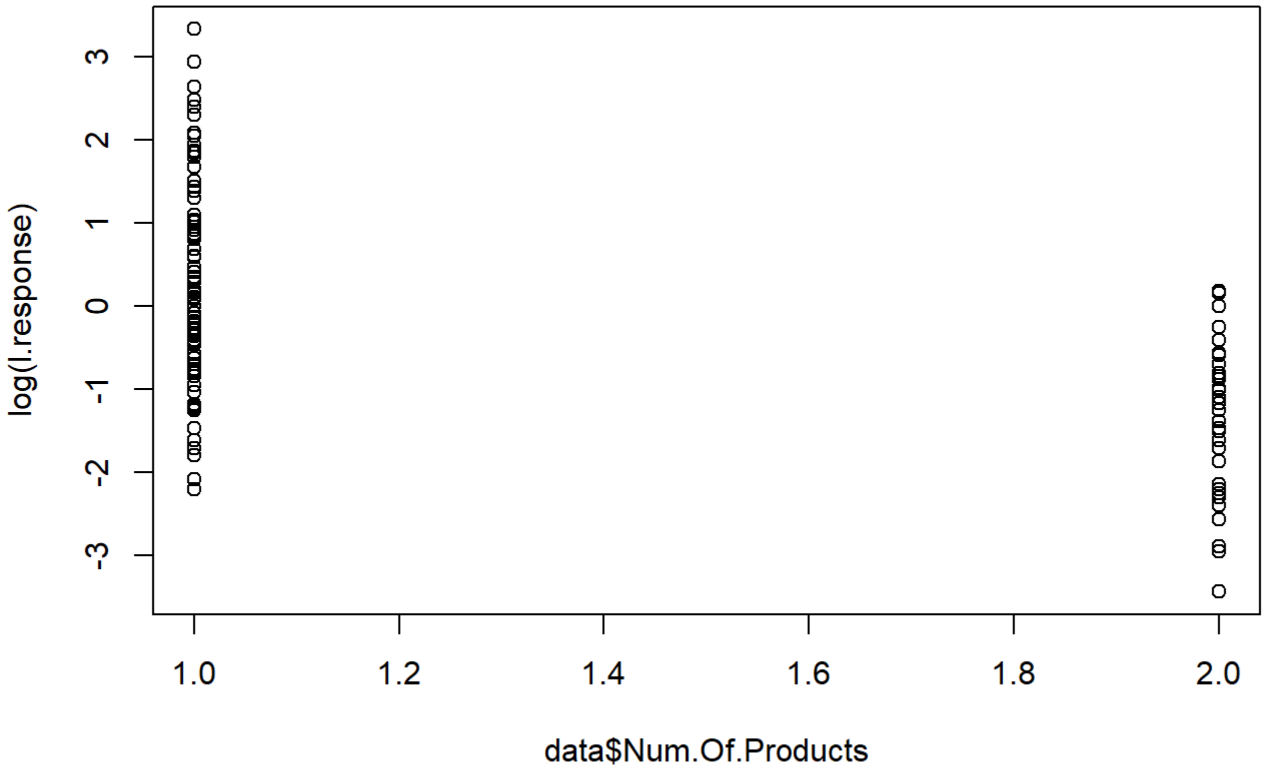


```
hist(res,xlab = "Std residuals", main ="",col= "orange")
```



Here, the predicting variable has only two values, so it is very hard to check for linearity assumption and also it is very hard to look for any pattern for Independence assumption.

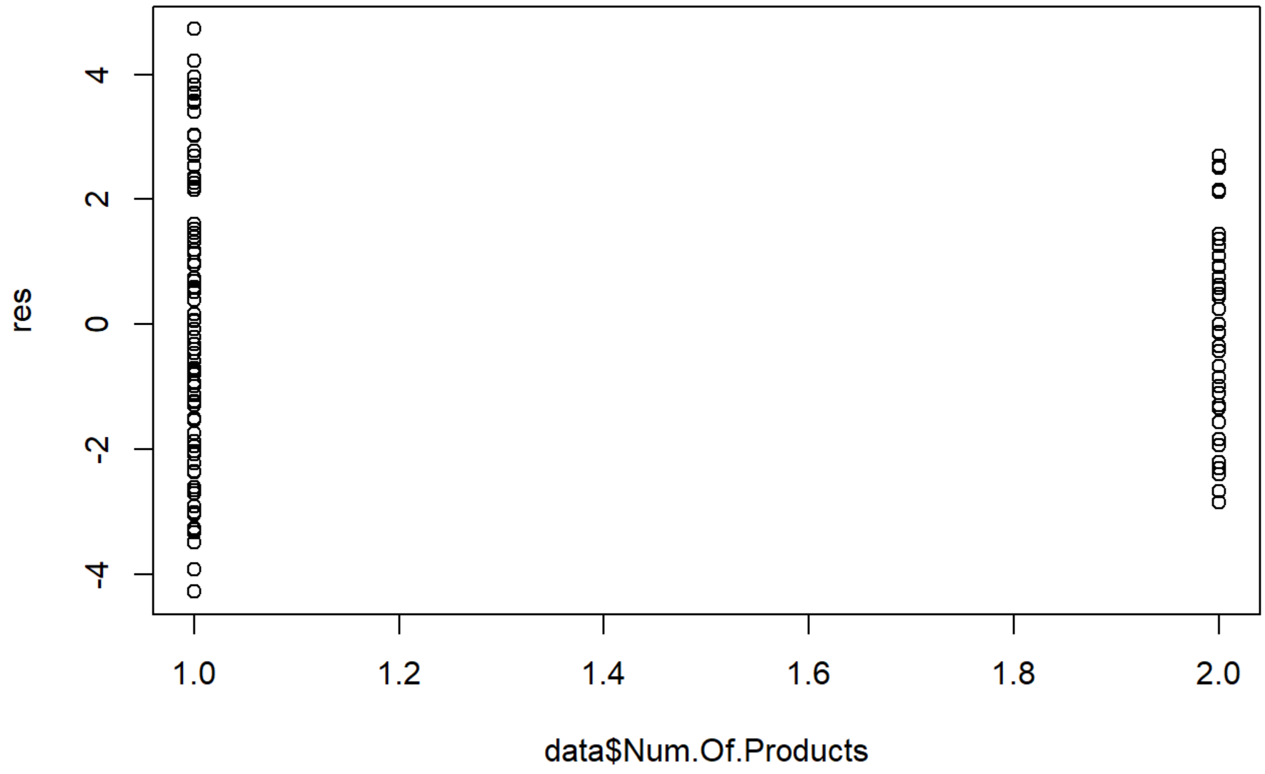
```
l.response <- (data$Staying/(1-data$Staying))
plot(data$Num.Of.Products, log(l.response))
```



I

think in below graph we don’t see any clusters, hence we can that the independe assumption holds.

```
plot(data$Num.Of.Products, res)
```



(c) 2 pts - Calculate the dispersion parameter for this model. Is this an overdispersed model? Yes this model is overdispersed as the estimate parameter is grater than 2.

```
sum(residuals(model1, type = "deviance")^2)/156
```

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

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 = Em
ployees,data= data, family= binomial)
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 of Staying = exp(Intercept +Age.Group + Gender + Tenure + Num.Of.Products + Is.Active.Member)/(1 + exp(Intercept +Age.Group + Gender + Tenure + Num.Of.Products + Is.Active.Member))

```
exp(-1.903330+ 1.229014-0.551438-0.003574-1.428767-0.871460)/(1+exp(-1.903330+ 1.229014-0.551438
-0.003574-1.428767-0.871460))
```

```
## [1] 0.0284829
```

(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.

The odds of staying increases by e^1.229014 = 3.418 unit with 1 unit increase in “Age.Group” keeping all other variables fixed.

The odds of staying decreases by 76% [1-e^(-1.428767) =1- 0.2396] with 1 unit increase in “Num.Of.Products” keeping all other variables fixed.

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

Yes, “Is.Active.Member” is very significant as it has very small p-value(2e-16), given the other variables.

(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.

From the below deviance and Pearson residuals we can accept the null hypothesis of good fit due to p-value being larger. Hence we can say that the model is a good fit.
Dispersion parameter is also less than 2 hence we can say that model is not overdispersed.

```
c(deviance(model2), 1-pchisq(deviance(model2),156))
```

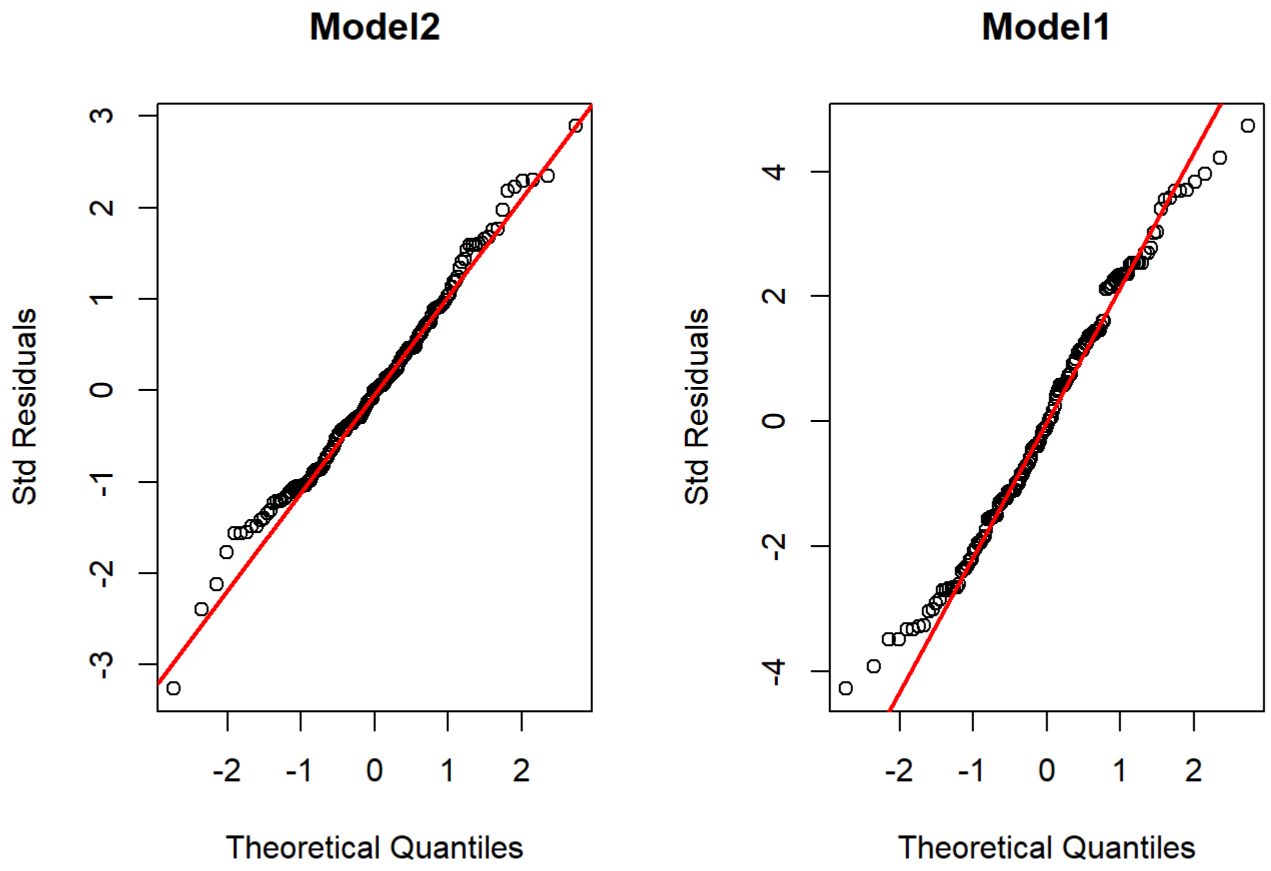
```
## [1] 171.9381966   0.1812144
```

```
pearson2 <- residuals(model2, type = "pearson")
pearson.tvalue2 <- sum(pearson2^2)
c(pearson.tvalue2, 1-pchisq(pearson.tvalue2,156))
```

```
## [1] 166.3908884   0.2698584
```

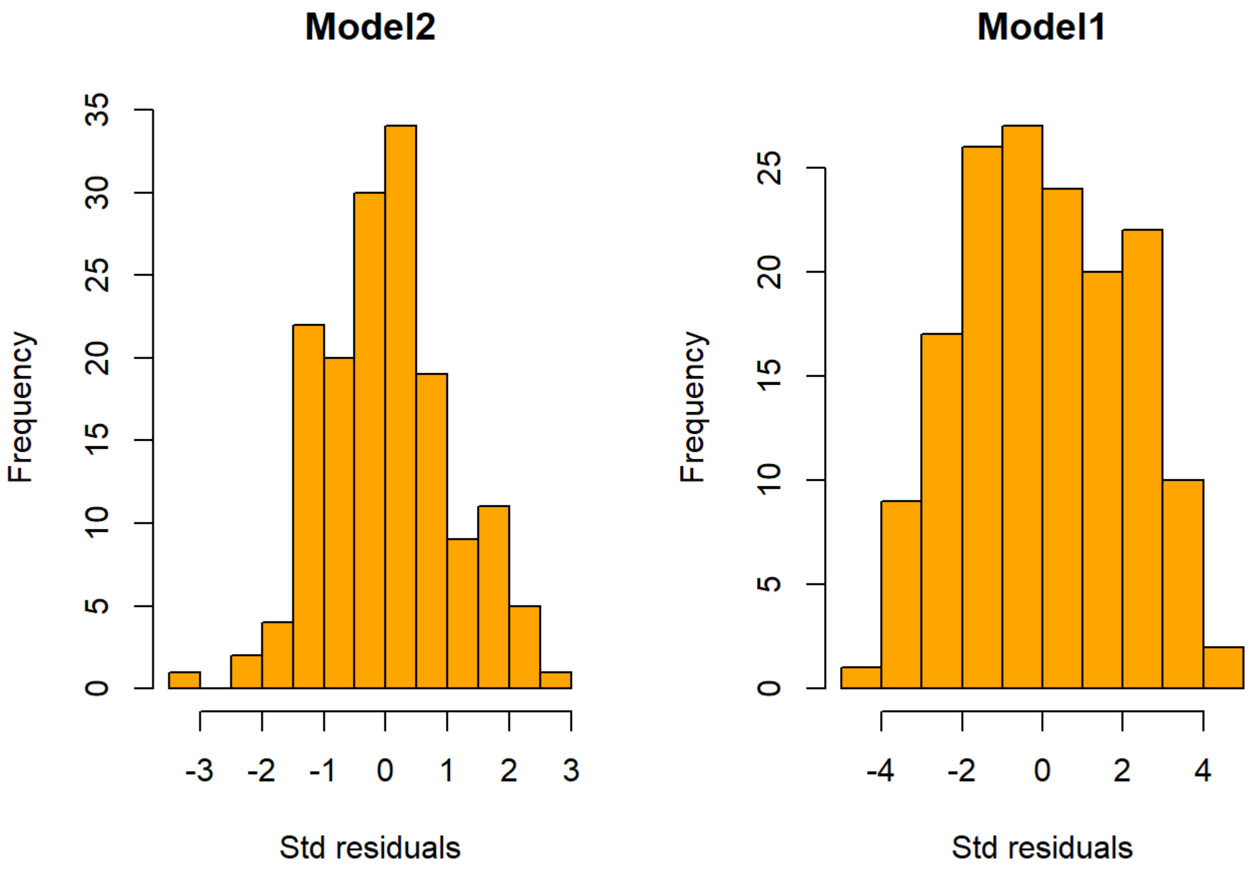
The below qq plot of model2 is little better than model1’s.

```
res2 <- resid(model2, type = "deviance")
par(mfrow = c(1,2))
qqnorm(res2, ylab = "Std Residuals", main ="Model2")
qqline(res2,col="red",lwd =2)
qqnorm(res, ylab = "Std Residuals", main = "Model1")
qqline(res,col="red",lwd =2)
```

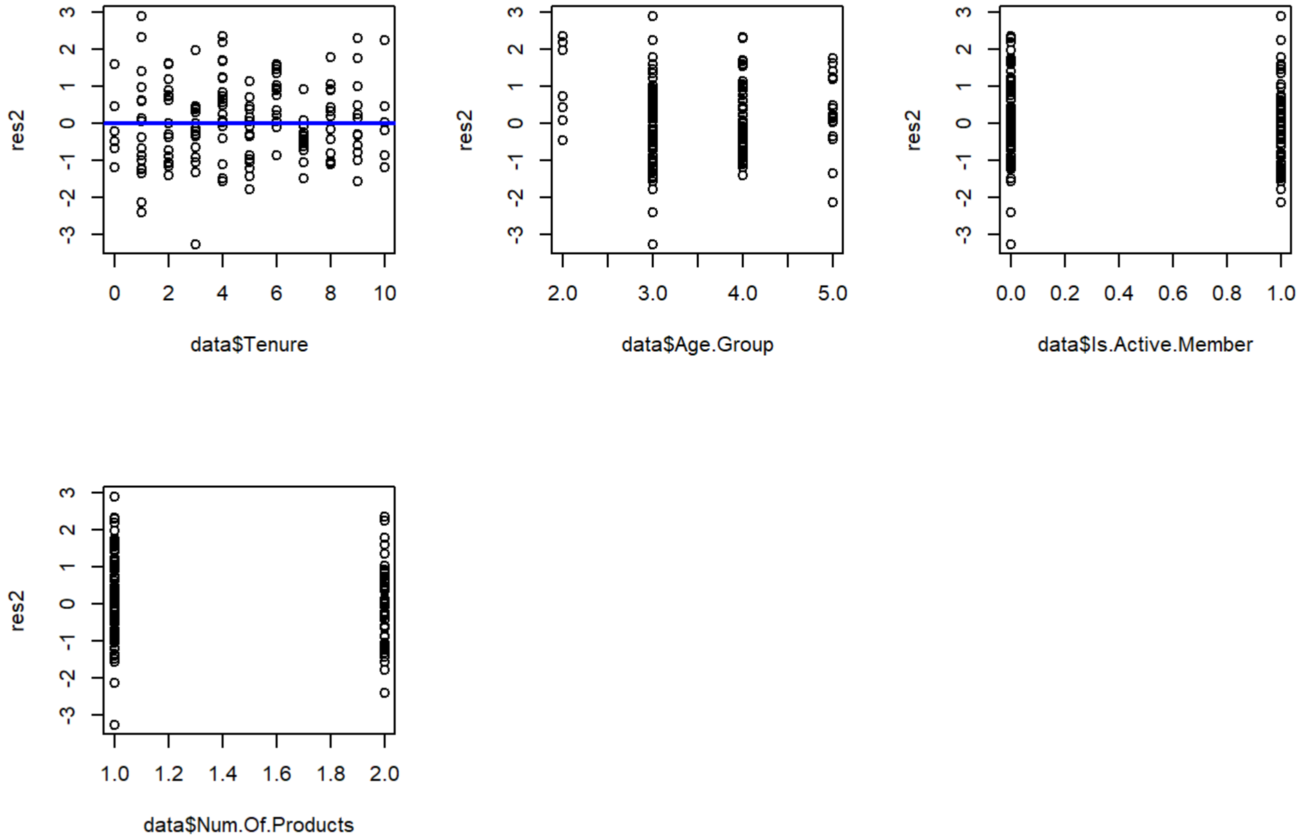


The model2's histogram is little better than of model1's.
Hence, we can say that in model2 the normality assumption holds little better than of model1's.

```
par(mfrow = c(1,2))
hist(res2,xlab = "Std residuals", main ="Model2",col= "orange")
hist(res,xlab = "Std residuals", main ="Model1",col= "orange")
```



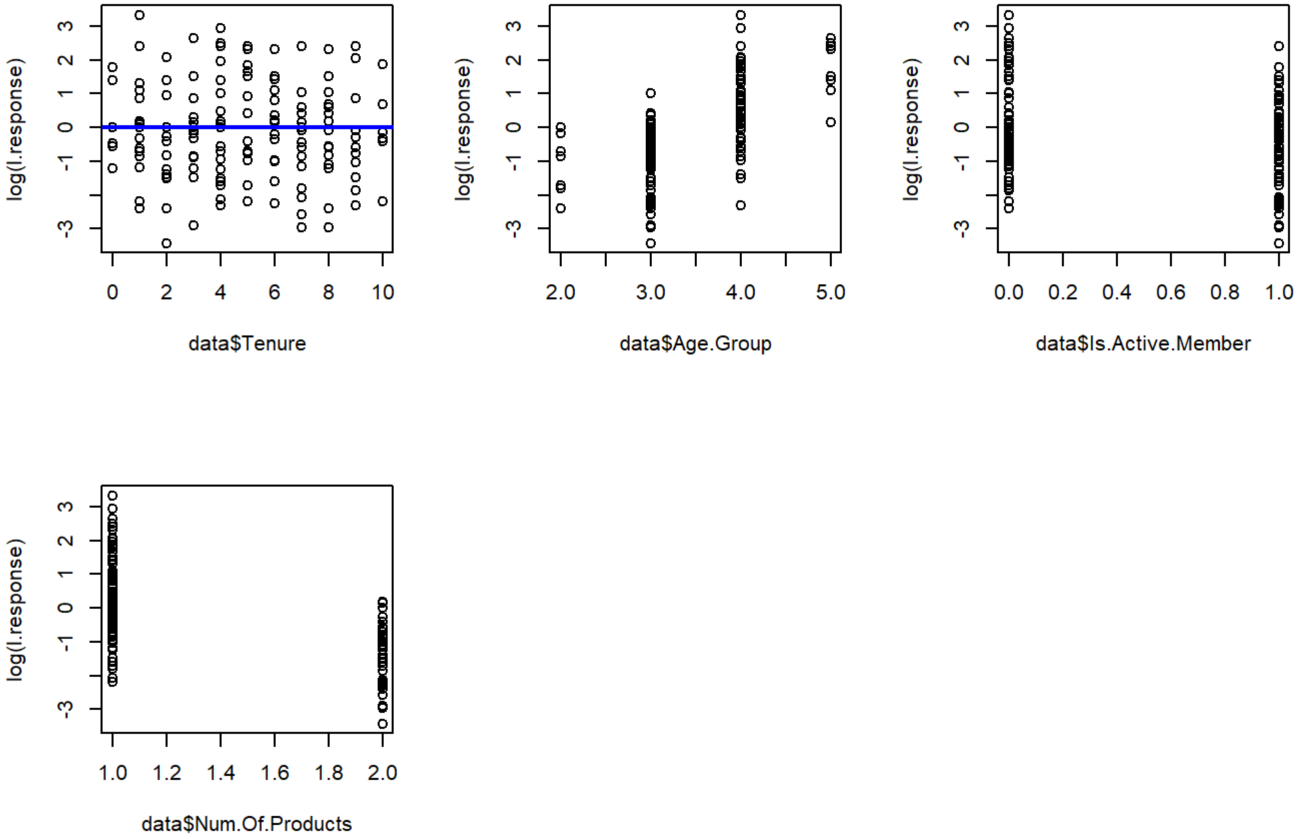
```
par(mfrow =c(2,3))
plot(data$Tenure,res2)
abline(0,0,col="blue",lwd=2)
plot(data$Age.Group,res2)
plot(data$Is.Active.Member,res2)
plot(data$Num.Of.Products,res2)
```



In the above graphs we don’t see any clusters of data for most of the predictors except the Age.Group, hence we can say that the independence assumption does not hold.

```
par(mfrow =c(2,3))

plot(data$Tenure,log(l.response))
abline(0,0,col="blue",lwd=2)
plot(data$Age.Group,log(l.response))
plot(data$Is.Active.Member,log(l.response))
plot(data$Num.Of.Products,log(l.response))
```



```
sum(residuals(model2, type = "deviance")^2)/152
```

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

From the above plot we can say that the linearity and independence assumption does not hold.

(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. First of all logistic regression works well with larger data set, Here we don't have large dataset. Keeping that in mind and based on the P-value for deviance and Pearson we can say that the model2 is a good-fitting model, however we can try other link function to check if that improves the goodness of fit. We can also add interaction terms to improve the goodness of fit. Below I have tried using "probit" link function which improves the goodness of fit by little margin.

```
model3 <- glm(Staying~ Age.Group + Gender+Tenure+Num.Of.Products +Is.Active.Member, weights = Em
ployees,data= data, family= binomial(link= "probit"))
summary(model3)
```

```
##
## Call:
## glm(formula = Staying ~ Age.Group + Gender + Tenure + Num.Of.Products +
##     Is.Active.Member, family = binomial(link = "probit"), data = data,
##     weights = Employees)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2640  -0.7560  -0.0259   0.6173   2.8162
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.117588   0.196933  -5.675 1.39e-08 ***
## Age.Group         0.729919   0.043128  16.925  < 2e-16 ***
## Gender           -0.331981   0.055058  -6.030 1.64e-09 ***
## Tenure           -0.001905   0.009781  -0.195    0.846
## Num.Of.Products  -0.851107   0.064427 -13.210  < 2e-16 ***
## Is.Active.Member -0.525565   0.055962  -9.392  < 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: 171.27  on 152  degrees of freedom
## AIC: 603.99
##
## Number of Fisher Scoring iterations: 4
```

```
c(deviance(model3), 1-pchisq(deviance(model3),156))
```

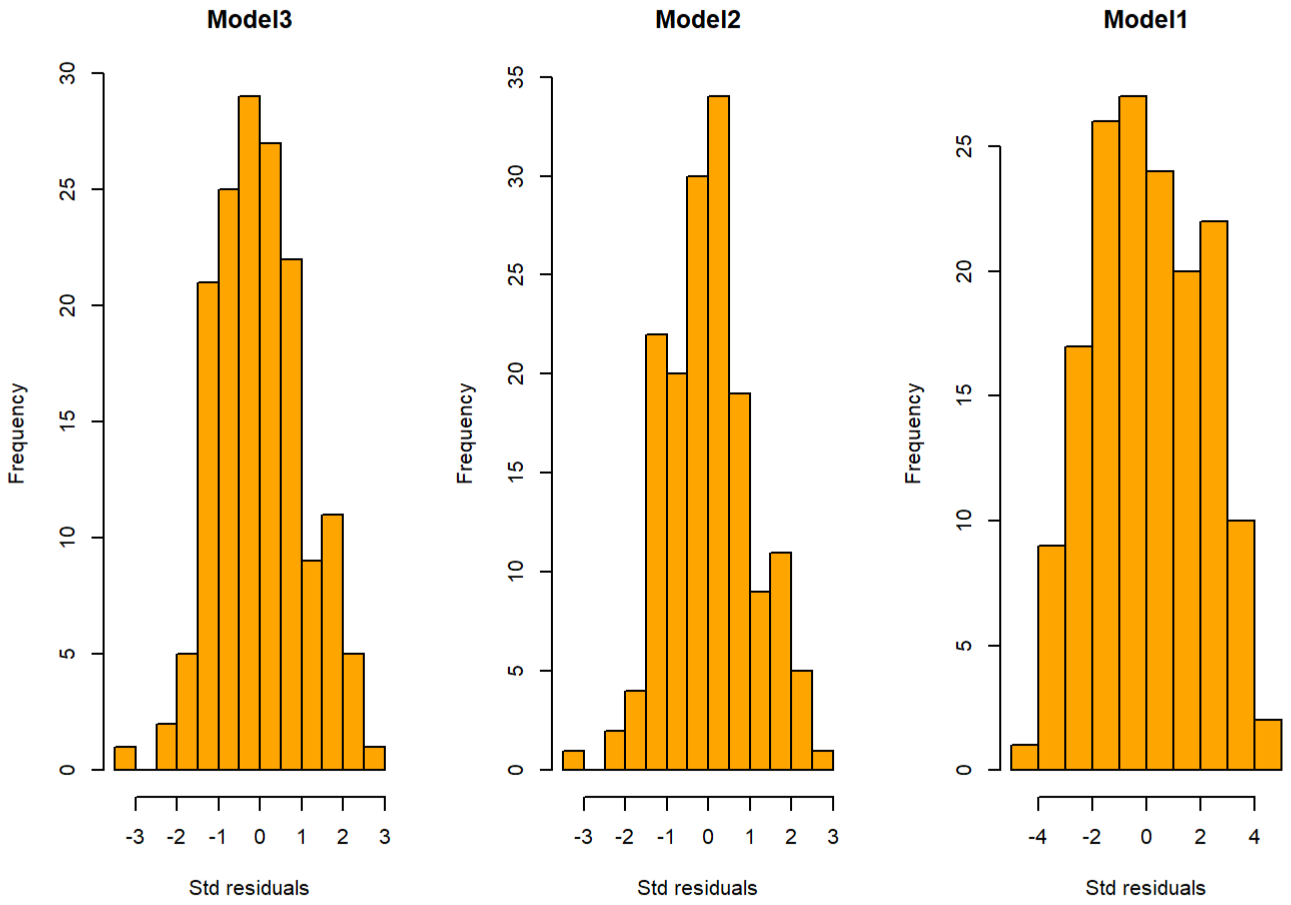
```
## [1] 171.2683305   0.1907429
```

```
pearson3 <- residuals(model3, type = "pearson")
pearson.tvalue3 <- sum(pearson3^2)
c(pearson.tvalue3, 1-pchisq(pearson.tvalue3,156))
```

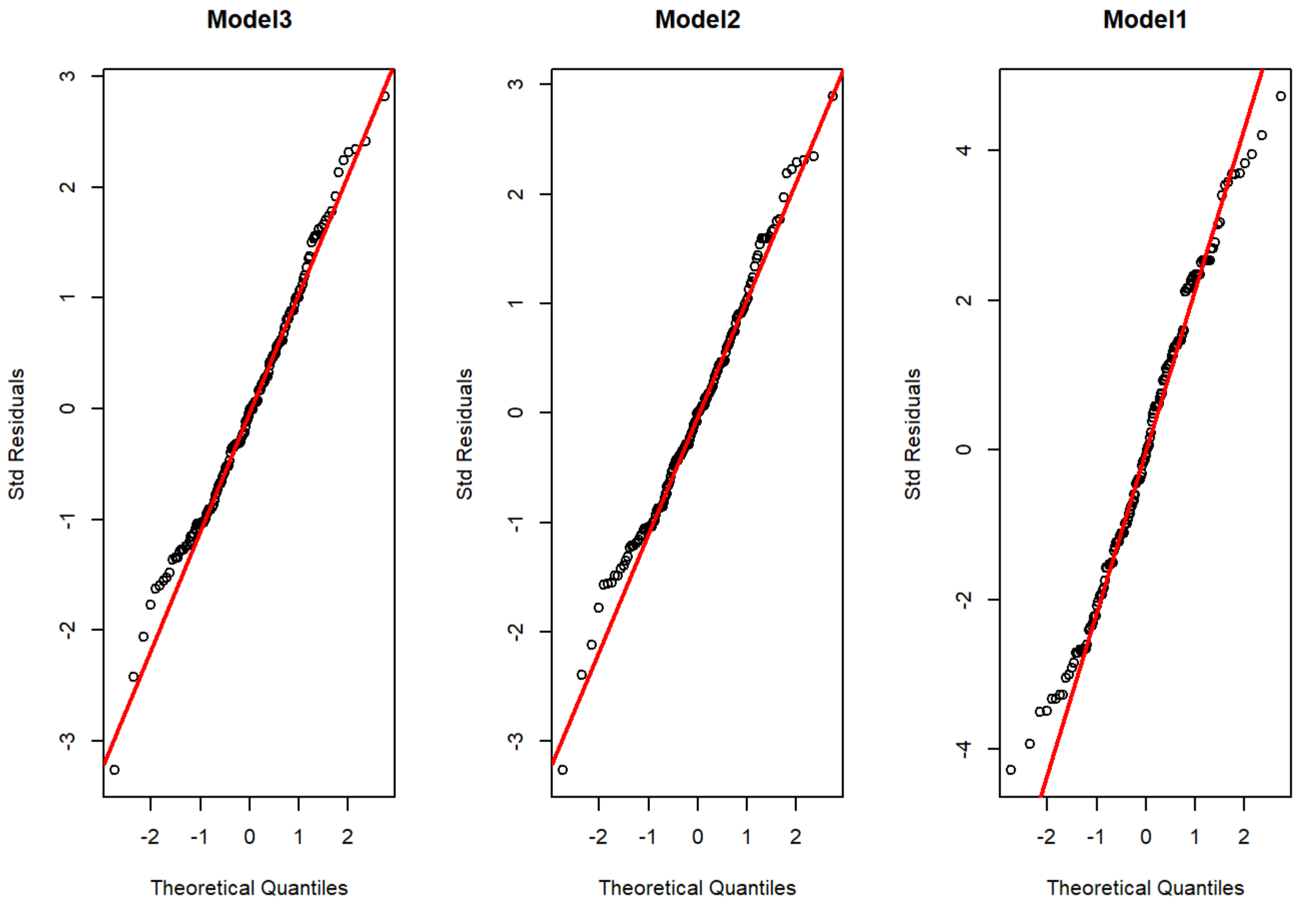
```
## [1] 166.2187977   0.2729525
```

```
res3 <-resid(model3, type = "deviance")
par(mfrow = c(1,3))

hist(res3,xlab = "Std residuals", main ="Model3",col= "orange")
hist(res2,xlab = "Std residuals", main ="Model2",col= "orange")
hist(res,xlab = "Std residuals", main ="Model1",col= "orange")
```



```
par(mfrow = c(1,3))
qqnorm(res3, ylab = "Std Residuals", main ="Model3")
qqline(res2,col="red",lwd =2)
qqnorm(res2, ylab = "Std Residuals", main ="Model2")
qqline(res2,col="red",lwd =2)
qqnorm(res, ylab = "Std Residuals", main = "Model1")
qqline(res,col="red",lwd =2)
```



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.

```
newdf <- data.frame(Num.Of.Products = 2)
m1 <- predict.glm(model1, newdf, type = "response")
m1
```

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

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

```
newdf2 <- data.frame(Age.Group =2,Gender =0,Tenure = 2, Is.Active.Member =1,Num.Of.Products =2)
m2 <- predict.glm(model2, newdf2, type = "response")
m2
```

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

(c) 2 pts - Comment on how your predictions compare. From the above result and given data model1 predicts the probability of employee staying as 0.1997 while the model2 predicts the probability of employee staying as 0.0498 which means that for the given data employee will not stay.