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
```

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, data = data, weights = Employees, 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
```

Model parameters are the intercept and the coefficient for Num.Of.Products. The estimate of the intercept is 2.1457and the estimate of the coefficient for Num.Of.Products is -1.7668.

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

$$\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 * Num.Of.Products} = e^{2.1457-1.7668 Num.Of.Products}$$

(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 a one unit increase in Num.Of.Products, the log-odds of staying will decrease by 1.7668, holding all other variables equal. A one unit increase in Num.Of.Products will change the odds of staying by a factor of 0.17088

$$e^{-1.7668} = 0.17088$$

Question 2: Inference - 9 pts

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

```
confidence <- confint.default(model1, level = 0.9)
confidence
```

```
##                         5 %      95 %
## (Intercept)       1.928820  2.362550
## Num.Of.Products -1.936459 -1.597197
```

The 90% confidence interval for the coefficient of Num.Of.Products is (-1.936, -1.597).

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

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

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

The p-value is 0 which indicates that at least one predicting variable significantly explains the fraction of employees that stay with the company (ie - Staying). In other words, the overall regression is significant.

(c) 3 pts - Which coefficients are significantly nonzero at the 0.01 significance level? Which are significantly negative? Why? Looking at the p values for the intercept and coefficient for Num.Of.Products, they are very close to 0 and are thus significant at the 0.01 significance level.

In order to test if the coefficient for Num.Of.Products is significantly negative, we check that the

$$\frac{pvalue}{2} < 0.01$$

. Since the p-value is 0, we can conclude that the coefficient for Num.Of.Products is indeed 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 Test
c(deviance(model1), 1 - pchisq(deviance(model1), 156))
```

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

```
# Pearson Residuals Test
pearson <- residuals(model1, type = "pearson")
pearson_tval <- sum(pearson^2)
c(pearson_tval, 1 - pchisq(pearson_tval, 156))
```

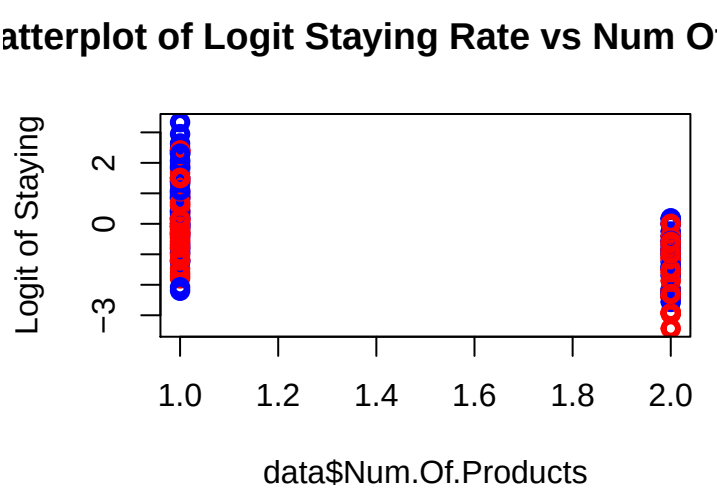
```
## [1] 562.1763   0.0000
```

The p values from the deviance test and Pearson residuals test are ~ 0 which indicates that we should reject the null hypothesis and that it is not a good fit. The results from 2b indicate that Num.Of.Products significantly explains the fraction of employees that stay with the company. The results do not contradict

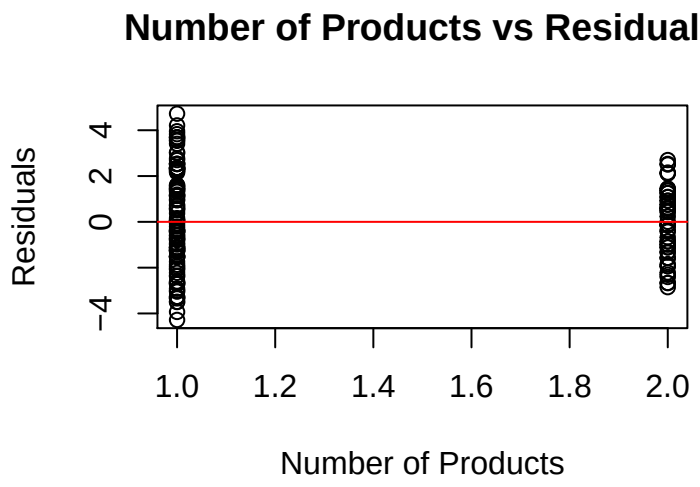
each other becaues they measure different things; the test done in 2b provides information on the predictive power of the model and the Deviance & Pearson Residual test provide information on the goodness of fit.

(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.

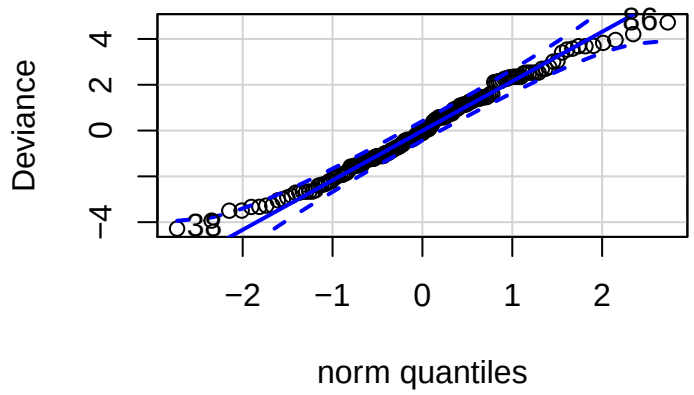
```
library("car")

# Logit of Staying vs Num of Products
plot(data$Num.Of.Products, log((data$Staying)/(1 - data$Staying)), ylab = "Logit of Staying",
    main = "Scatterplot of Logit Staying Rate vs Num Of Products", col = c("red",
        "blue"), lwd = 3)
```




```
# Num of Products vs Residuals
res <- resid(model1, type = "deviance")
plot(data$Num.Of.Products, res, xlab = "Number of Products", ylab = "Residuals",
    main = "Number of Products vs Residuals")
abline(0, 0, col = "red")
```

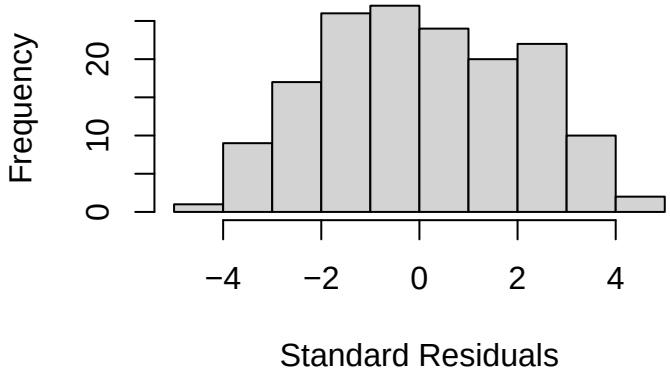



```
# QQ plot and histogram
qqPlot(res, ylab = "Deviance", main = "")
```

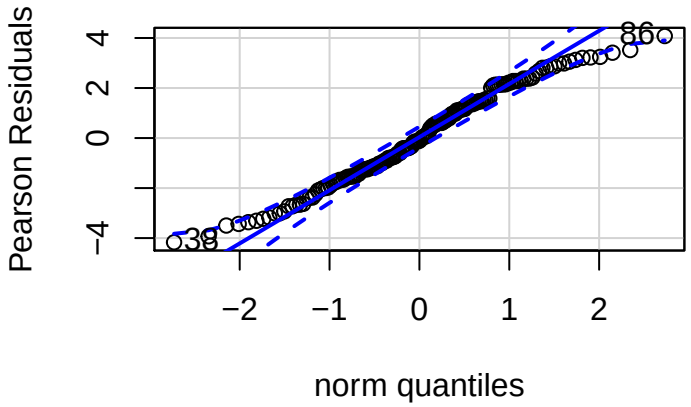



```
## [1] 86 38
```

```
hist(res, 10, xlab = "Standard Residuals", main = "")
```

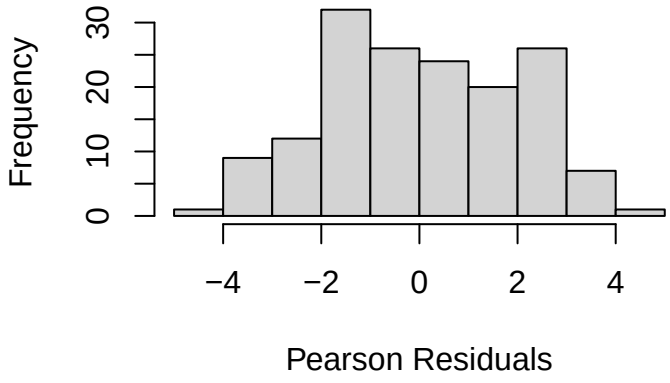



```
qqPlot(pearson, ylab = "Pearson Residuals", main = "")
```




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

```
hist(pearson, 10, xlab = "Pearson Residuals", main = "")
```



1) Linearity - does not hold; using the Logit of Staying vs Num of Products plot, we can see that linearity is violated as Num of Products only takes on 2 values.

2) Independence - holds; using the Num of Product vs Residuals plot, we can see that there is no clear pattern or clustering of residuals.

3) Residuals - holds; the deviance residuals and the Pearson residuals look approximately normal.

Despite the linearity assumption being violated, the plots indicate that model1 may be a good fit for the data.

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

```
D <- sum((residuals(model1, type = "deviance")^2))
phi <- D/(158 - 1 - 1)
phi
```

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

$$\phi = 4.05$$

. This indicates that the model is overdispersed.

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,
    data = data, weights = Employees, 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.

$$p(x) = \frac{e^{\beta_0+\beta_1 Age.Group+\beta_2 Gender+\beta_3 Tenure+\beta_4 Num.Of.Products+\beta_5 Is.Active.Member}}{1+e^{\beta_0+\beta_1 Age.Group+\beta_2 Gender+\beta_3 Tenure+\beta_4 Num.Of.Products+\beta_5 Is.Active.Member}}$$

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

(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.

For a one unit increase in Age.Group, the log-odds of Staying will increase by 1.229014, holding all other variables equal. A one unit increase in Age.Group will change the odds of staying by a factor of 3.41786, holding all variables equal.

$$e^{1.229014} = 3.41786$$

For a one unit increase in Is.Active.Member, the log-odds of Staying will decrease by 0.871460, holding all other variables equal. A one unit increase in Is.Active.Member will change the odds of staying by a factor of 0.41834, holding all other variables equal.

$$e^{-0.871460} = 0.41834$$

(c) 2.5 pts - Is *Is.Active.Member* significant given the other variables in model2? The p value for Is.Active.Member is ~ 0, indicating that it is a statistically significant predictor.

(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.

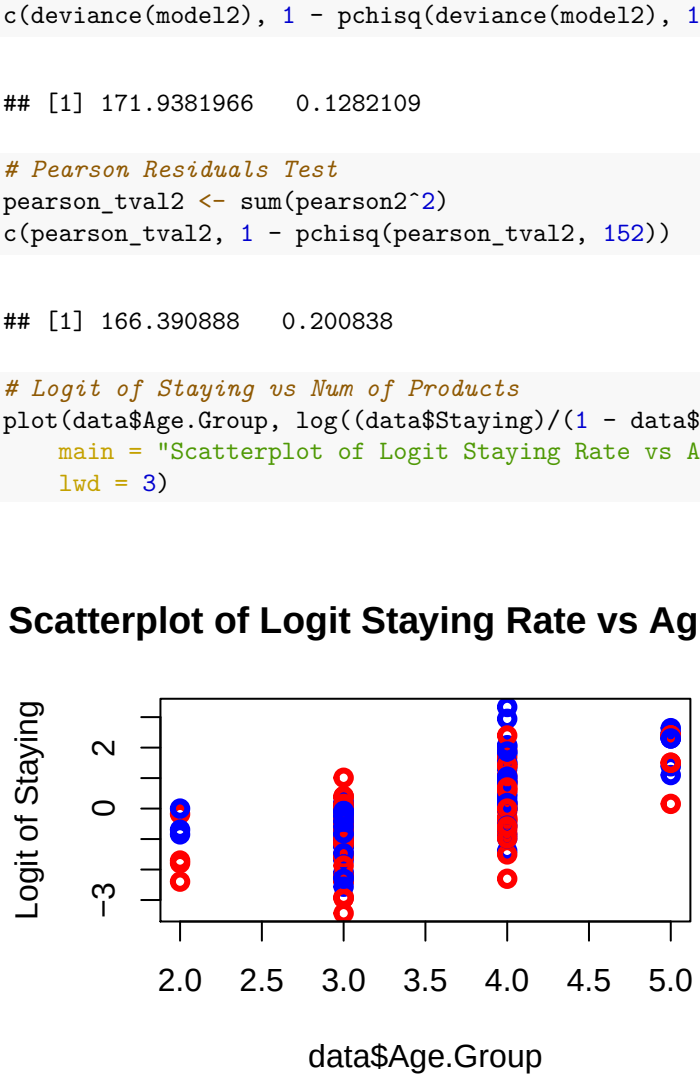
```r
res2 <- residuals(model2, type = "deviance")
pearson2 <- residuals(model2, type = "pearson")

# Deviance Test
c(deviance(model2), 1 - pchisq(deviance(model2), 152))
```

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

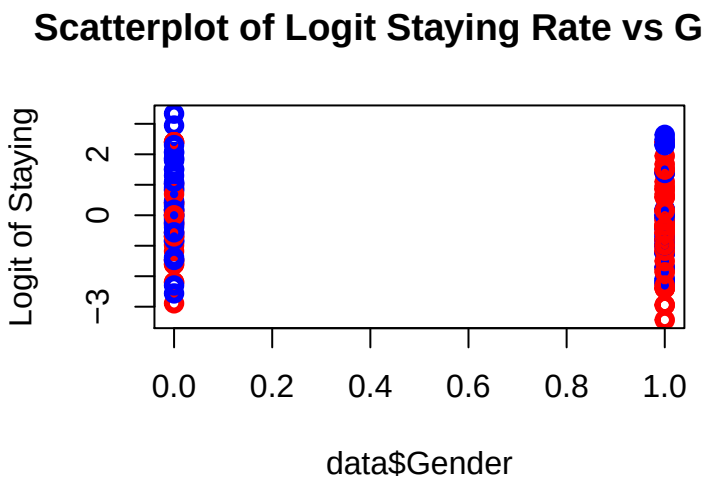
```r
# Pearson Residuals Test
pearson_tval2 <- sum(pearson2^2)
c(pearson_tval2, 1 - pchisq(pearson_tval2, 152))
```

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

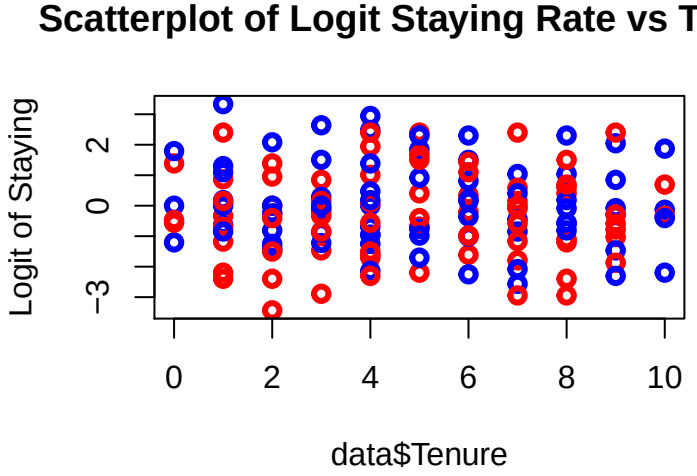
```r
# Logit of Staying vs Num of Products
plot(data$Age.Group, log((data$Staying)/(1 - data$Staying)), ylab = "Logit of Staying",
    main = "Scatterplot of Logit Staying Rate vs Age.Group", col = c("red", "blue"),
    lwd = 3)
```




```r
plot(data$Gender, log((data$Staying)/(1 - data$Staying)), ylab = "Logit of Staying",
    main = "Scatterplot of Logit Staying Rate vs Gender", col = c("red", "blue"),
    lwd = 3)
```

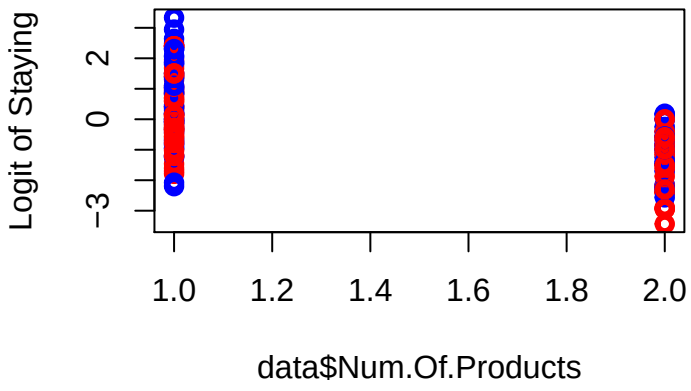



```
plot(data$Tenure, log((data$Staying)/(1 - data$Staying)), ylab = "Logit of Staying",
    main = "Scatterplot of Logit Staying Rate vs Tenure", col = c("red", "blue"),
    lwd = 3)
```



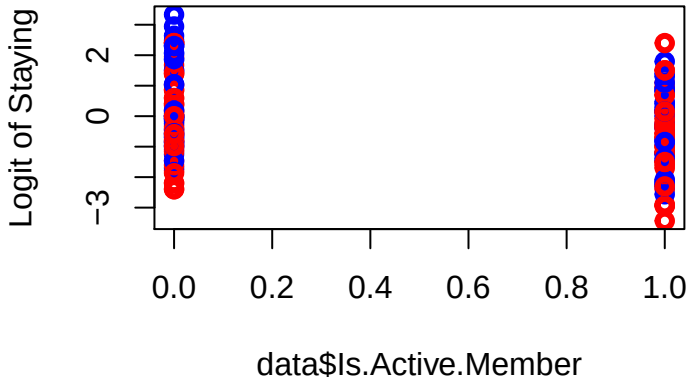

```
plot(data$Num.Of.Products, log((data$Staying)/(1 - data$Staying)), ylab = "Logit of Staying",
    main = "Scatterplot of Logit Staying Rate vs Num Of Products", col = c("red",
        "blue"), lwd = 3)
```

atterplot of Logit Staying Rate vs Num Of Pr

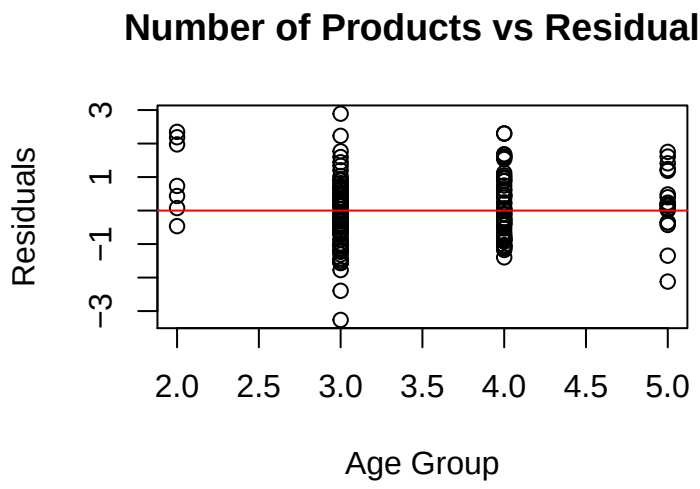



```
plot(data$Is.Active.Member, log((data$Staying)/(1 - data$Staying)), ylab = "Logit of Staying",
    main = "Scatterplot of Logit Staying Rate vs Is Active Member", col = c("red",
        "blue"), lwd = 3)
```

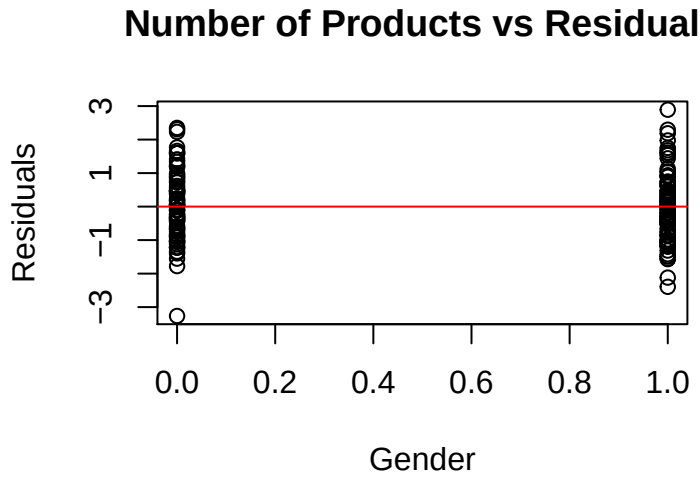
atterplot of Logit Staying Rate vs Is Active M



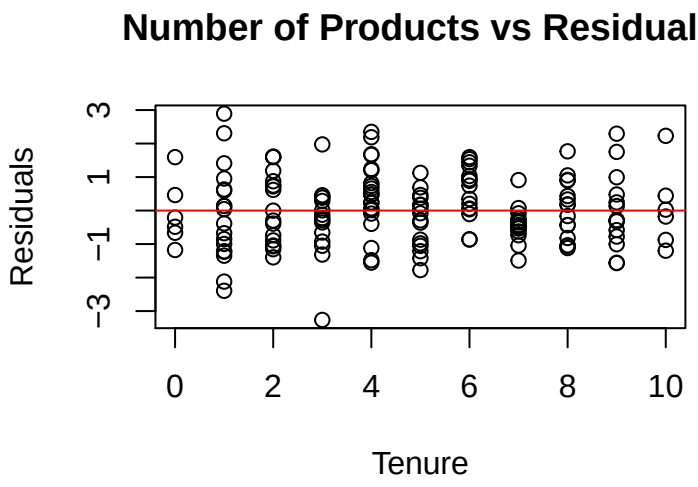

```
# Predictors vs Residuals
plot(data$Age.Group, res2, xlab = "Age Group", ylab = "Residuals", main = "Number of Products vs Residua
abline(0, 0, col = "red")
```



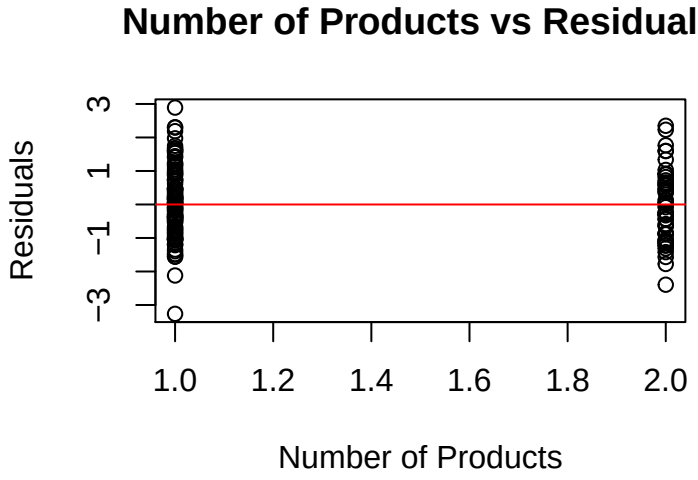

```
plot(data$Gender, res2, xlab = "Gender", ylab = "Residuals", main = "Number of Products vs Residuals")
abline(0, 0, col = "red")
```



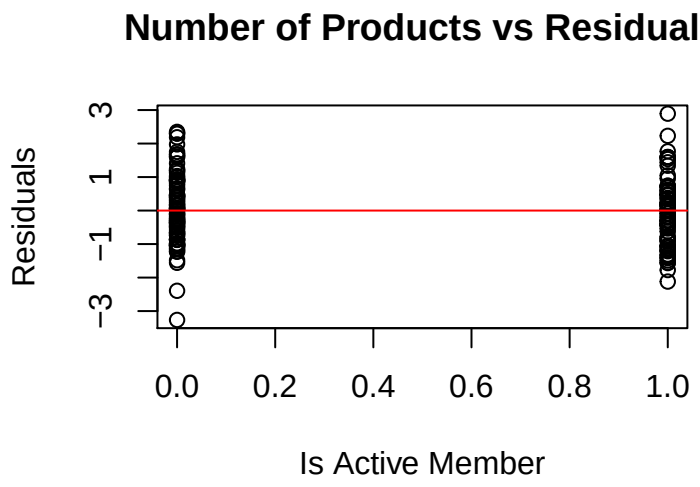

```
plot(data$Tenure, res2, xlab = "Tenure", ylab = "Residuals", main = "Number of Products vs Residuals")
abline(0, 0, col = "red")
```



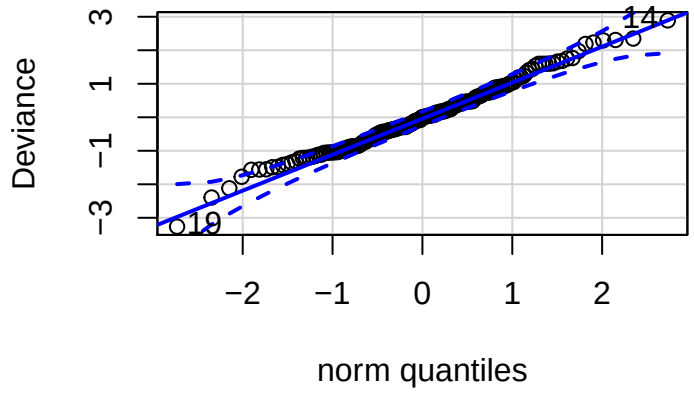

```
plot(data$Num.Of.Products, res2, xlab = "Number of Products", ylab = "Residuals",
    main = "Number of Products vs Residuals")
abline(0, 0, col = "red")
```




```
plot(data$Is.Active.Member, res2, xlab = "Is Active Member", ylab = "Residuals",
    main = "Number of Products vs Residuals")
abline(0, 0, col = "red")
```

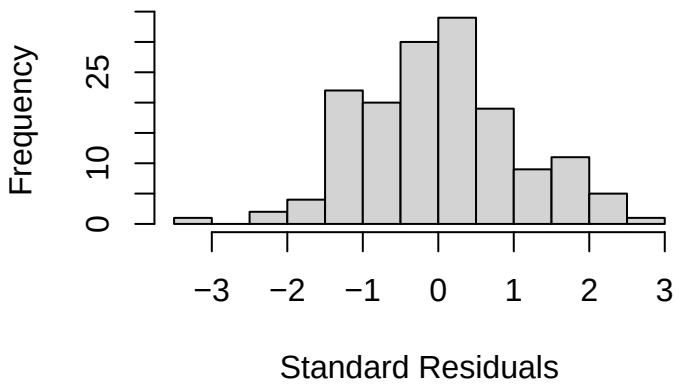



```
# QQ plot and histogram
qqPlot(res2, ylab = "Deviance", main = "")
```

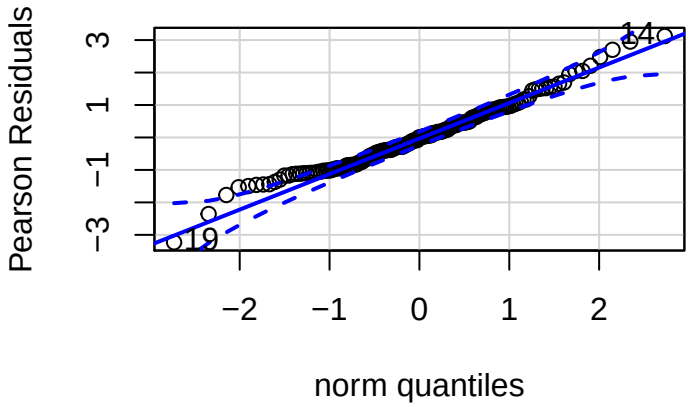



```
## [1] 19 14
```

```
hist(res2, 10, xlab = "Standard Residuals", main = "")
```

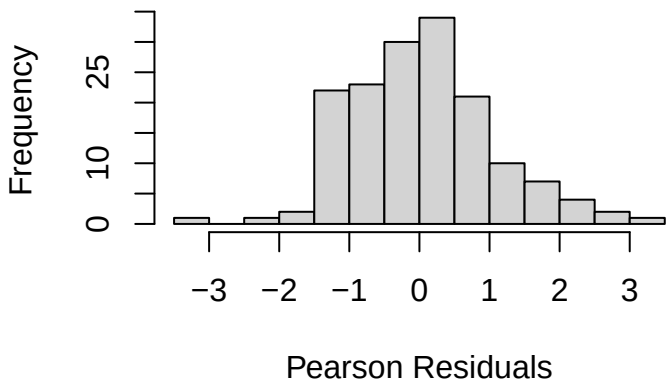



```
qqPlot(pearson2, ylab = "Pearson Residuals", main = "")
```




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

```
hist(pearson2, 10, xlab = "Pearson Residuals", main = "")
```




```
# Dispersion
D2 <- sum(res2^2)
phi2 <- D2/(158 - 5 - 1)
phi2
```

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

The p values for the Deviance and Pearson Residuals tests are now large, suggesting that we can fail to reject the null hypothesis and that the model may be a good fit.

Checking the model assumptions:
1) Linearity - does not hold; using the Logit of Staying vs Predictor plots, we can see that linearity cannot be practically assessed because most predictors take on only a few values. The plot for Age.Group shows an increasing trend for the logit value as Age.Group increases. 2) Independence - holds; using the Predictor vs Residuals plots, we can see that there is no clear pattern or clustering of residuals for any of the predictors. 3) Residuals - holds; the deviance residuals and the Pearson residuals look approximately normal.

Despite the linearity assumption being violated, the plots indicate that model2 may also be a good fit for the data.

Dispersion for model2 is 1.13, indicating that there is no more overdispersion in this model.

(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.
Overall, I believe that model2 is better fitting model compared to model1. An improvement could be done to change the link function or transform the data or to analyze the outliers of the data to get the linearity assumption to hold.

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.

```
new_data <- data.frame(Age.Group = 2, Gender = 0, Tenure = 2, Num.Of.Products = 2,
    Is.Active.Member = 1)
predict(model1, new_data, type = "response")
```

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

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

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

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

(c) 2 pts - Comment on how your predictions compare.
When taking Age.Group, Gender, Tenure and Is.Active.Member into account, the estimate of the probability of staying decreases from 20% to 4%. Based on the analysis performed above, model2 may be more reliable than model1 as it takes more factors into consideration and is a better fit model.