Generalized Linear Models

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.

Fitting Model

A logistic regression model using *Staying* as the response variable with *Num.Of.Products* as the predictor and logit as the link function was fit while ensuring to include the weights parameter for specifying the number of trials. The summary of this model can be seen below:

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

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

For this model, the equation for the Odds of Staying is:

$$\frac{p(Staying)}{1-p(Staying)} = e^{\beta_0+\beta_1 x} = e^{0.37886-1.76683*Num.Of.Products2}$$

Coefficient Interpretation

The estimated coefficients for *Num.Of.Products2* is associated with the log odds of employees who stayed with the company who also owned two products. Therefore, A 1-unit increase in the number of products owned decreases the log odds of employees staying with the company by 1.76683. Additionally, a 1-unit increase in the number of products owned decreases the odds of employees staying with the company by 82.91% ($1 - e^{-1.7668} = 0.8291$). A 90% confidence interval for the coefficients for *Num.Of.Products2* can be seen below:

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

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

Model Interpretation

As seen previously in the model summary, the null deviance associated with model1 is 981.04 on 157 degrees of freedom, and the residual deviance is 632.04 on 156 degrees of freedom. Given this information, the p-value associated with the test for overall significance can be seen below:

```
1-pchisq(981.04-632.04, 157-156)
```

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

The p-value is approximately 0, therefore using a significance level of 0.01 we would conclude that the model is significant.

As seen previously in the model summary, the p-values associated with the test for coefficients being non-zero were 1.37e-15 and < 2e-16 for β_0 and β_1 respectively. Given that both of these p-values are less than the significance level of 0.01, we would conclude that they are significantly nonzero. Additionally, the test for being significantly negative requires only half of each of these p-values, which would still be less than 0.01 (one-tail test rather than two). Given that the coefficient estimate for β_1 is negative and the associated p-value is less thatn the significance level, we can assume that β_1 is significantly negative.

Goodness of Fit

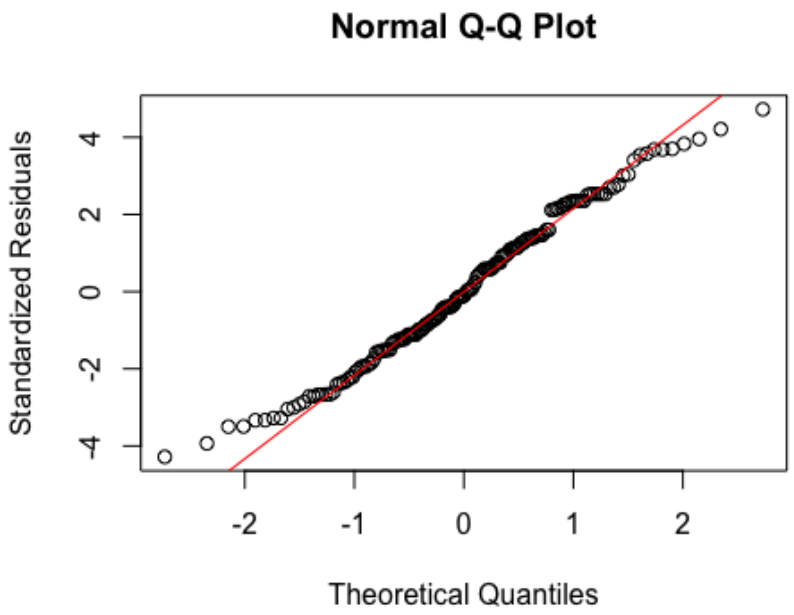
The goodness-of-fit hypothesis tests for both the Deviance and Pearson residuals on 156 degrees of freedom can be seen above. Both of these test have p-values which are approximately 0. Given that the null hypothesis for this test is that the logistic regression model fits the data, we would reject the null hypothesis and conclude that the model does not fit the data.

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

## [1] 0

pear_res = residuals(model1,type="pearson")
pear_t = sum(pear_res^2)
1-pchisq(pear_t,156)

## [1] 0
```

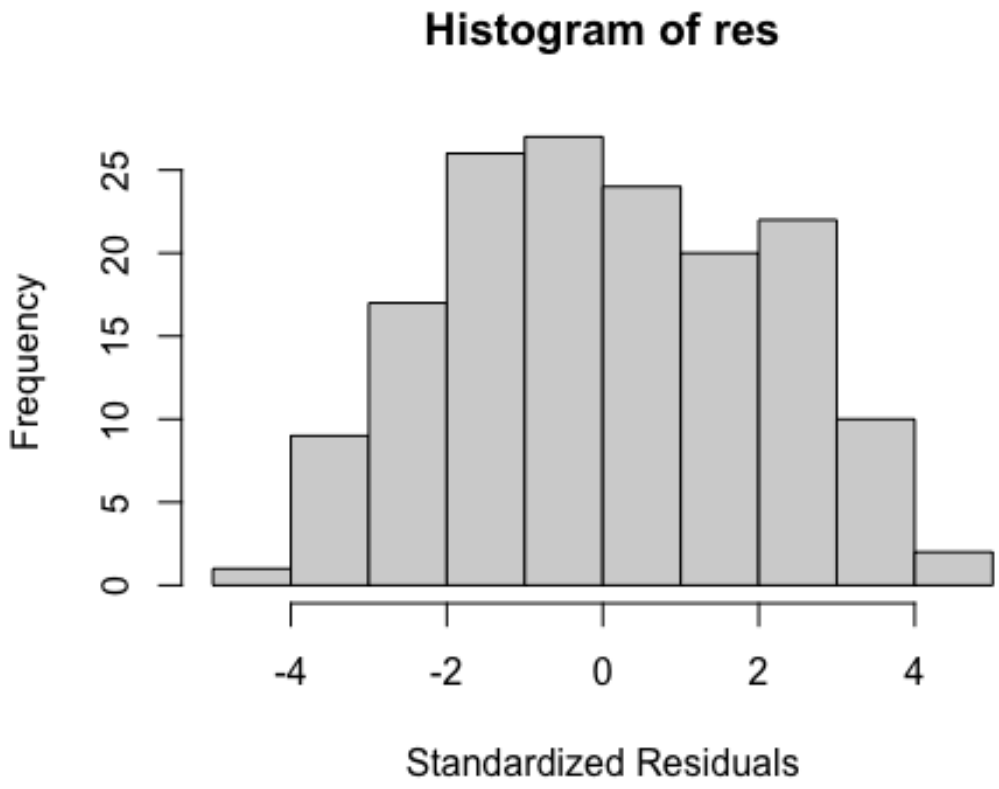
Previously it was concluded that the overall model was significant, which seems slightly contradictory since here it was concluded that the model does not fit that data. It is important to note, however, that it is possible for a model to have predictive power even if it is a poor fit. Therefore, the results found here are not contrary to those found previously.

Both the QQ plot and the histogram seen above show a fairly normal distribution of the standardized deviance residuals, which suggests that the normality assumption is met.

```
res = resid(model1, type='deviance')
qqnorm(res, ylab='Standardized Residuals')
qqline(res, col='red', lwd=1)
```



```
hist(res, xlab='Standardized Residuals')
```



The estimated dispersion parameter for this model is greater than 2, which would suggest that the model is overdispersed.

```
model1$deviance/model1$df.res
```

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

Fitting a Full Model

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 was fit while ensuring to include the weights parameter for specifying the number of trials. The summary for this model can be seen below:

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

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

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

The equation for the probability of staying is:

$$p(x) = \frac{e^{-0.109572+0.384480*Age.Group3+1.734115*Age.Group4+2.955578*Age.Group5-0.572069*Gender1-0.003319*Tenure-1.410}}{1+e^{-0.109572+0.384480*Age.Group3+1.734115*Age.Group4+2.955578*Age.Group5-0.572069*Gender1-0.003319*Tenure-1.4}}$$

Coefficient Interpretation

A 1-unit increase in Tenure decreases the odds of employees staying with the company by 0.33%, while holding all other predictors constant ($1 - e^{-0.003319} = 0.003313498$).

Being an active member decreases the odds of employees staying with the company by 57.27% compared to non-active members, while holding all other predictors constant ($1 - e^{-0.850280} = 0.5727047$).

Given the summary of the model seen previously, it can be seen that the p-value associated with the test for significance for *Is.Active.Member1* is < 2e-16. Given that this is less than 0.01, we can conclude that *Is.Active.Member1* in this model, given the other variables in the model.

Goodness of Fit

The goodness-of-fit hypothesis tests for both the Deviance and Pearson residuals on 150 degrees of freedom can be seen above. Both of these test have p-values which are greater than 0.01. Given that the null hypothesis for this test is that the logistic regression model fits the data, we would fail to reject the null hypothesis and conclude that the model fits the data.

```
1-pchisq(deviance(model2),150)
```

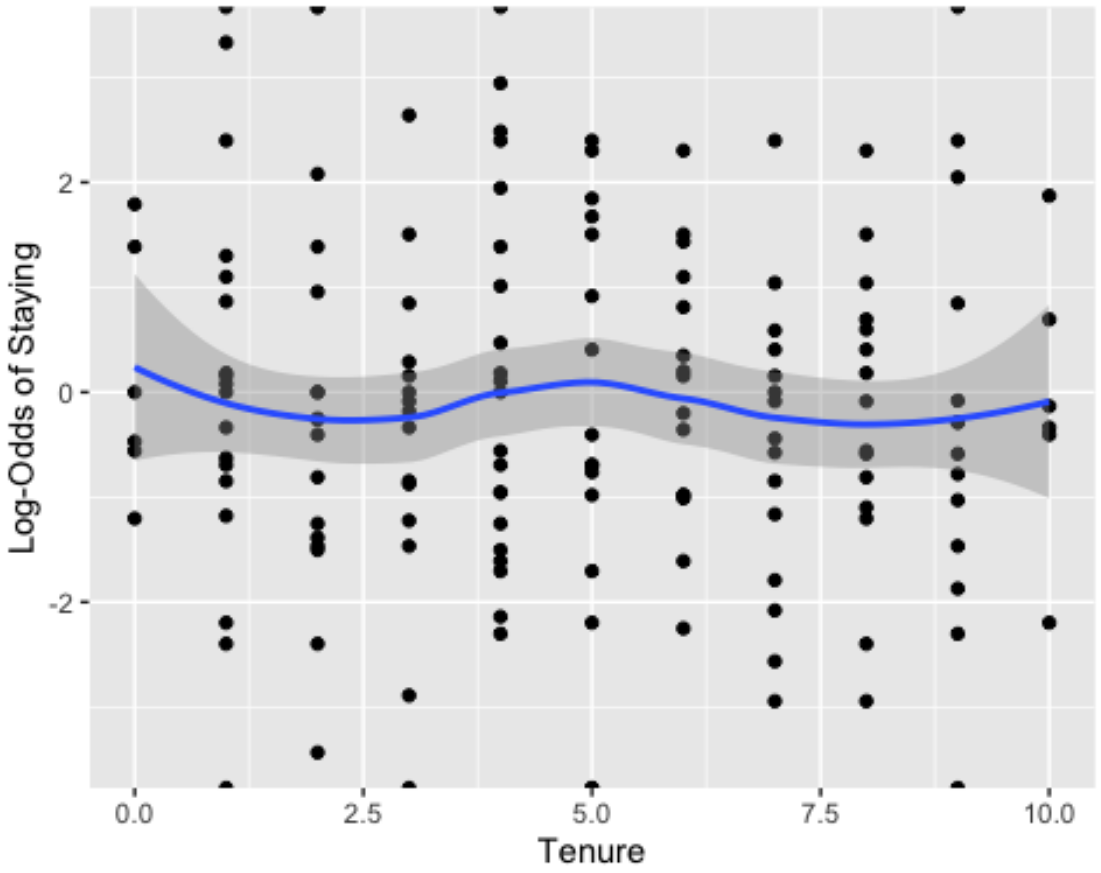
```
## [1] 0.2319118
```

```
pear_res = residuals(model2,type="pearson")
pear_t = sum(pear_res^2)
1-pchisq(pear_t,150)
```

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

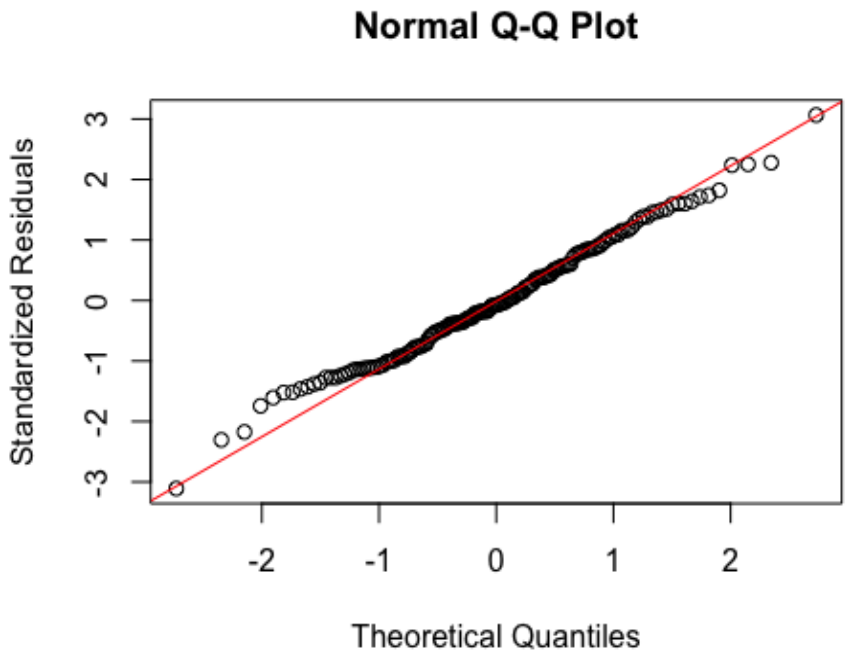
The relationship between log-odds of Staying and Tenure does not appear to have a strong linear relationship, or much of a relationship at all.

```
ggplot(rawdata, aes(Tenure, log((Staying)/(1-Staying))) ) +
  geom_point() +
  stat_smooth() +
  ylab('Log-Odds of Staying')
```

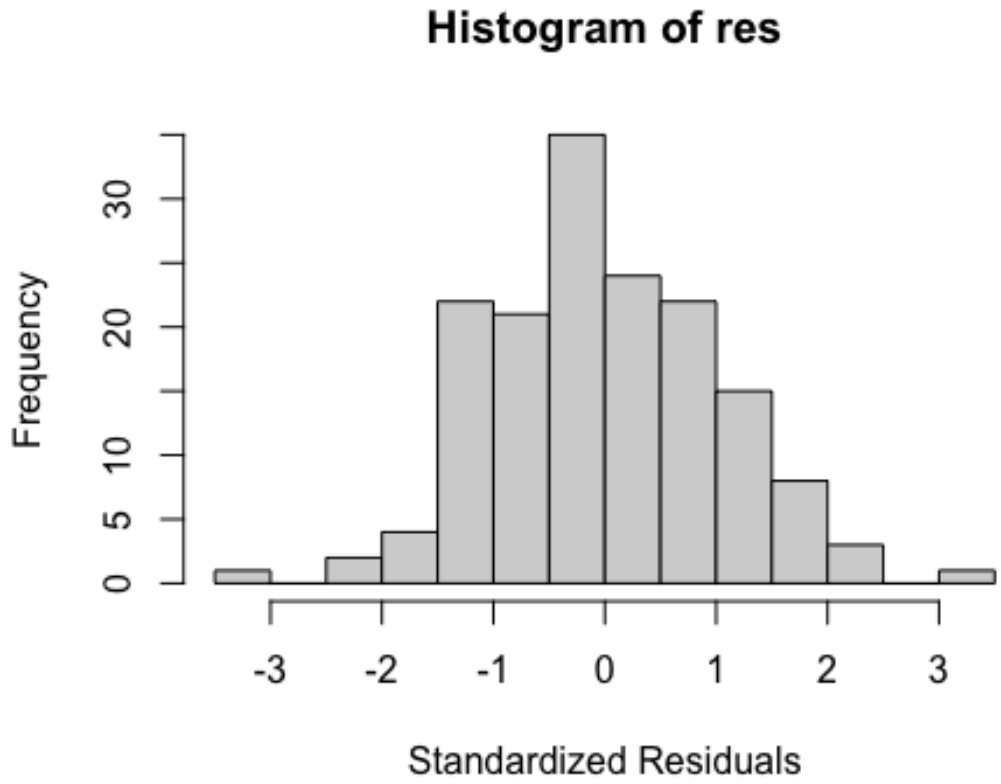


Both the QQ plot and the histogram seen above show a fairly normal distribution of the standardized deviance residuals, which suggests that the normality assumption is met.

```
res = resid(model2, type='deviance')
qqnorm(res, ylab='Standardized Residuals')
qqline(res, col='red', lwd=1)
```




```
hist(res, xlab='Standardized Residuals')
```



The estimated dispersion parameter for this model is less than zero and is approximately 1, therefore we could conclude that this model is not overdispersed.

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

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

Given that the model assumptions seem to have improved from model1 to model2 and it was concluded that model2 fits the data whereas model1 did not, it can be concluded that model2 is a good fitting model.

A possible improvement to this model could be to test different link functions to see if one have a lower deviance and therefore possible a better fit. Additionally, removing possible outliers could improve the fit of the model, but it should be noted that these possible outliers should be examined carefully before being removed.