Titanic_Survival

Laura Cebollero

22/10/2018

Dataset description

We first load the data and check it:

```
nrow(TitanicSurvival)
```

```
## [1] 1309
```

```
kable(summary(TitanicSurvival))
```

survived	sex	age	passengerClass
no :809	female:466	Min. : 0.1667	1st:323
yes:500	male :843	1st Qu.:21.0000	2nd:277
NA	NA	Median :28.0000	3rd:709
NA	NA	Mean :29.8811	NA
NA	NA	3rd Qu.:39.0000	NA
NA	NA	Max. :80.0000	NA
NA	NA	NA's :263	NA

There are 4 variables:

- Survived: Boolean. Whether the passenger survived or not.
- Sex: Categorical. 2 categories. The gender of the passenger.
- Age: Numerical. The age of the passenger.
- PassengerClass: Categorical. 3 categories. The ticket class of the passenger.

Exploratory data analysis

263 observations that have missing age. We remove these obs. because we are interested in using age onwards.

```
df = as.data.frame(TitanicSurvival)

# Tables
# Of NA ONLY!!! To see where are more NAs
df2 = df[is.na(df$age), ]
table(df2$sex)
```

```
##
## female   male
##     78    185
```

```
table(df2$passengerClass)
```

```
##
## 1st 2nd 3rd
##  39  16 208
```

```r
table(df2$passengerClass, df2$sex)
```

```
##
##       female male
##   1st     11   28
##   2nd      3   13
##   3rd     64  144
```

```r
t(prop.table(table(df2$passengerClass, df2$sex)))
```

```
##
##                 1st        2nd        3rd
##   female 0.04182510 0.01140684 0.24334601
##   male   0.10646388 0.04942966 0.54752852
```

```r
# We could remove them, but we do not want to. Better to impute.
# df = df[!is.na(df$age), ]
#
# We impute with the avg. cell of each cell.
kable(aggregate(age ~ sex + passengerClass, data = df[!is.na(df), ], FUN = mean))
```

sex	passengerClass	age
female	1st	37.03759
male	1st	41.02925
female	2nd	27.49919
male	2nd	30.81540
female	3rd	22.18531
male	3rd	25.96227

```r
res = (aggregate(age ~ sex + passengerClass, data = df[!is.na(df), ], FUN = mean))
res = as.data.frame(res)
res
```

```
##      sex passengerClass      age
## 1 female            1st 37.03759
## 2   male            1st 41.02925
## 3 female            2nd 27.49919
## 4   male            2nd 30.81540
## 5 female            3rd 22.18531
## 6   male            3rd 25.96227
```

```r
for (x in seq(1, length(res$age))) {
    r = res[x, ]
    for(j in seq(1, length(df$age))){
        row = df[j, ]
        if(is.na(row$age) & row$sex == r$sex & row$passengerClass == r$passengerClass){
            df[j, 'age'] <- r$age
        }
    }
}
```

We want our response variable to be whether a passenger survived or not depending on the variables sex, age and passengerClass. And what about their interaction?

Model definition & analysis

Model 1

```r
model1 <- glm(survived ~ sex + passengerClass + age, data = df, family=binomial)
deviance(model1)
```

```
## [1] 1225.889
```

```r
AIC(model1)
```

```
## [1] 1235.889
```

```r
sum(residuals(model1, type = "pearson")^2)  # Pearson test. X^2
```

```
## [1] 1354.965
```

```r
sum(residuals(model1, type = "pearson")^2)/(nrow(df) - 5)   # X^2/(N-P)
```

```
## [1] 1.039083
```

```r
# Close to 1. Empirical dispersion. We are doing well assuming a binomial distribution with dispersion p
#
# Remember:
# AIC = -2l + 2p ;;; Where p = num of parameters
# AIC useful to penalize models with large number of parameters

# With AIC we cannot perform an Hypothesis test.
# To compare, we can use X^2 (chi square)
```

```r
model1 <- glm(survived ~ sex + passengerClass + age, data = df, family=binomial)
deviance(model1)
```

```
## [1] 1225.889
```

```r
AIC(model1)
```

```
## [1] 1235.889
```

```r
# p = 5
sum(residuals(model1, type = "pearson")^2)  # Pearson test. X^2
```

```
## [1] 1354.965
```

```r
sum(residuals(model1, type = "pearson")^2)/(nrow(df) - 5)   # X^2/(N-P)
```

```
## [1] 1.039083
```

```r
# Close to 1. Empirical dispersion. We are doing well assuming a binomial distribution with dispersion p
#
# Remember:
# AIC = -2l + 2p ;;; Where p = num of parameters
# AIC useful to penalize models with large number of parameters

# With AIC we cannot perform an Hypothesis test.
# To compare, we can use X^2 (chi square)
```

```r
summary(model1)
```

```
##
```

```
## Call:
## glm(formula = survived ~ sex + passengerClass + age, family = binomial,
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6139  -0.6851  -0.4616   0.6724   2.4063
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        3.451335   0.311561  11.078  < 2e-16 ***
## sexmale           -2.465359   0.148483 -16.604  < 2e-16 ***
## passengerClass2nd -1.256703   0.214038  -5.871 4.32e-09 ***
## passengerClass3rd -2.281515   0.206034 -11.073  < 2e-16 ***
## age               -0.034282   0.006288  -5.452 4.99e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1741.0  on 1308  degrees of freedom
## Residual deviance: 1225.9  on 1304  degrees of freedom
## AIC: 1235.9
##
## Number of Fisher Scoring iterations: 4
```

All variables are significant (***). Female taken as baseline. -2.46 for male.

If we have a man with:

- same age as a woman
- same class

THEN it has a lower prob. to survive than the woman, for its ODDS ratio = $e^{-2.46}$

```r
Anova(model1)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: survived
##                LR Chisq Df Pr(>Chisq)
## sex              329.48  1  < 2.2e-16 ***
## passengerClass   142.19  2  < 2.2e-16 ***
## age               31.33  1  2.174e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
Anova(model1, ty=3)
```

```
## Analysis of Deviance Table (Type III tests)
##
## Response: survived
##                LR Chisq Df Pr(>Chisq)
## sex              329.48  1  < 2.2e-16 ***
## passengerClass   142.19  2  < 2.2e-16 ***
## age               31.33  1  2.174e-08 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is no interaction. Type 2 and Type 3 result in the same in the tests.

Tests globally to check if the variables are significant in the model.

Model 2

Let us put more interactions.

Let's see if the 2 categorical values interact:

```r
model2 <- glm(survived ~ (sex * passengerClass) + age, data = df, family=binomial)

summary(model2)
```

```
##
## Call:
## glm(formula = survived ~ (sex * passengerClass) + age, family = binomial,
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1112  -0.6253  -0.5225   0.4394   2.4989
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 4.91023    0.54610   8.991  < 2e-16 ***
## sexmale                    -4.00376    0.48671  -8.226  < 2e-16 ***
## passengerClass2nd          -1.68162    0.55888  -3.009  0.00262 **
## passengerClass3rd          -4.08146    0.50076  -8.151 3.62e-16 ***
## age                        -0.03915    0.00678  -5.774 7.74e-09 ***
## sexmale:passengerClass2nd   0.12497    0.61706   0.203  0.83951
## sexmale:passengerClass3rd   2.42492    0.52240   4.642 3.45e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1741.0  on 1308  degrees of freedom
## Residual deviance: 1174.4  on 1302  degrees of freedom
## AIC: 1188.4
##
## Number of Fisher Scoring iterations: 6
```

```r
# p = 7. AIC penalizes this!!!

deviance(model2)
```

```
## [1] 1174.381
```

```r
AIC(model2)
```

```
## [1] 1188.381
```

```r
sum(residuals(model2, type = "pearson")^2)  # Pearson test. X^2
```

```
## [1] 1397.005
```

```r
sum(residuals(model2, type = "pearson")^2)/(nrow(df) - length(model2$coefficients))   # X^2/(N-P)
```

```
## [1] 1.072968
```

Based that the AIC is lower although it penalizes the num of parameters, we choose this model because it supposedly fits more.

Since Model1 is nested in model2, we can compare the deviances:

```r
AIC(model1) - AIC(model2)
```

```
## [1] 47.50868
```

With 2 degrees of freedom. Which is larger than $\chi^2_{0.05,2}$.

```r
qchisq(0.95, 1)
```

```
## [1] 3.841459
```

So we reject the null hypothesis and we prefer model2 instead of model1.

Model 3

```r
model3 <- glm(survived ~ (sex * passengerClass) + (age * sex), data = df, family=binomial)

summary(model3)
```

```
##
## Call:
## glm(formula = survived ~ (sex * passengerClass) + (age * sex),
##     family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8534  -0.6355  -0.5047   0.4690   2.6446
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 4.09344    0.64198   6.376 1.81e-10 ***
## sexmale                    -2.76798    0.74273  -3.727 0.000194 ***
## passengerClass2nd          -1.46586    0.56215  -2.608 0.009118 **
## passengerClass3rd          -3.69148    0.51809  -7.125 1.04e-12 ***
## age                        -0.01981    0.01116  -1.775 0.075833 .
## sexmale:passengerClass2nd -0.23164    0.63711  -0.364 0.716173
## sexmale:passengerClass3rd  1.86684    0.57465   3.249 0.001159 **
## sexmale:age                -0.03009    0.01412  -2.132 0.033036 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1741.0  on 1308  degrees of freedom
## Residual deviance: 1169.9  on 1301  degrees of freedom
## AIC: 1185.9
##
## Number of Fisher Scoring iterations: 5
```

```r
# p = 7. AIC penalizes this!!!

deviance(model3)
```

```
## [1] 1169.863
```

```r
AIC(model3)
```

```
## [1] 1185.863
```

```r
sum(residuals(model3, type = "pearson")^2)  # Pearson test. X^2
```

```
## [1] 1342.535
```

```r
sum(residuals(model3, type = "pearson")^2)/(nrow(df) - length(model3$coefficients))   # X^2/(N-P)
```

```
## [1] 1.031925
```

To check that the age*sex is significant:

```r
Anova(model3, ty=3)
```

```
## Analysis of Deviance Table (Type III tests)
##
## Response: survived
##                     LR Chisq Df Pr(>Chisq)
## sex                   15.994  1  6.355e-05 ***
## passengerClass       119.767  2  < 2.2e-16 ***
## age                    3.196  1    0.07380 .
## sex:passengerClass    32.398  2  9.221e-08 ***
## sex:age                4.518  1    0.03355 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It is!

Model3 seems better than model2!

Model 4

```r
model4 <- glm(survived ~ (sex * passengerClass)  + (age * passengerClass), data = df, family=binomial)

summary(model4)
```

```
##
## Call:
## glm(formula = survived ~ (sex * passengerClass) + (age * passengerClass),
##     family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0120  -0.6276  -0.5134   0.3632   3.0386
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  4.588758   0.676343   6.785 1.16e-11 ***
## sexmale                     -3.973904   0.485723  -8.181 2.80e-16 ***
```

```
## passengerClass2nd          0.044423   0.957224   0.046  0.96298
## passengerClass3rd         -4.021435   0.722655  -5.565 2.62e-08 ***
## age                       -0.031721   0.011695  -2.712  0.00668 **
## sexmale:passengerClass2nd -0.293664   0.661713  -0.444  0.65719
## sexmale:passengerClass3rd  2.373134   0.520669   4.558 5.17e-06 ***
## passengerClass2nd:age     -0.048474   0.020230  -2.396  0.01657 *
## passengerClass3rd:age      0.004426   0.015196   0.291  0.77082
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1741.0  on 1308  degrees of freedom
## Residual deviance: 1165.7  on 1300  degrees of freedom
## AIC: 1183.7
##
## Number of Fisher Scoring iterations: 6
```

```r
# p = 7. AIC penalizes this!!!

deviance(model4)
```

```
## [1] 1165.658
```

```r
AIC(model4)
```

```
## [1] 1183.658
```

```r
sum(residuals(model4, type = "pearson")^2)  # Pearson test. X^2
```

```
## [1] 1452.841
```

```r
sum(residuals(model4, type = "pearson")^2)/(nrow(df) - length(model4$coefficients))   # X^2/(N-P)
```

```
## [1] 1.11757
```

Based on the deviance, we may reject model2 for the diff is greater the Chisq:

```r
AIC(model2) - AIC(model4)
```

```
## [1] 4.722459
```

```r
qchisq(0.95, 2)
```

```
## [1] 5.991465
```

```r
AIC(model2) - AIC(model4)  > qchisq(0.95, 2)
```

```
## [1] FALSE
```

For now, model3 is still the winner.

Model 5

```r
model5 <- glm(survived ~ (sex * passengerClass) + (age * sex) + (age * passengerClass), data = df, famil

summary(model5)
```

```
##
```

```
## Call:
## glm(formula = survived ~ (sex * passengerClass) + (age * sex) +
##     (age * passengerClass), family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6247  -0.6416  -0.4828   0.4007   3.2023
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  3.41707    0.80737   4.232 2.31e-05 ***
## sexmale                     -2.63277    0.74979  -3.511 0.000446 ***
## passengerClass2nd            0.50273    0.93419   0.538 0.590479
## passengerClass3rd           -3.21663    0.76305  -4.215 2.49e-05 ***
## age                         -0.00247    0.01780  -0.139 0.889628
## sexmale:passengerClass2nd -0.61314    0.66414  -0.923 0.355900
## sexmale:passengerClass3rd  1.80245    0.56553   3.187 0.001437 **
## sexmale:age                 -0.03356    0.01521  -2.207 0.027343 *
## passengerClass2nd:age       -0.05745    0.02099  -2.736 0.006217 **
## passengerClass3rd:age       -0.00824    0.01646  -0.501 0.616532
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1741.0  on 1308  degrees of freedom
## Residual deviance: 1160.8  on 1299  degrees of freedom
## AIC: 1180.8
##
## Number of Fisher Scoring iterations: 5
```

```r
# p = 7. AIC penalizes this!!!

deviance(model5)
```

```
## [1] 1160.773
```

```r
AIC(model5)
```

```
## [1] 1180.773
```

```r
sum(residuals(model5, type = "pearson")^2)  # Pearson test. X^2
```

```
## [1] 1462.958
```

```r
sum(residuals(model5, type = "pearson")^2)/(nrow(df)  - length(model5$coefficients))   # X^2/(N-P)
```

```
## [1] 1.126219
```

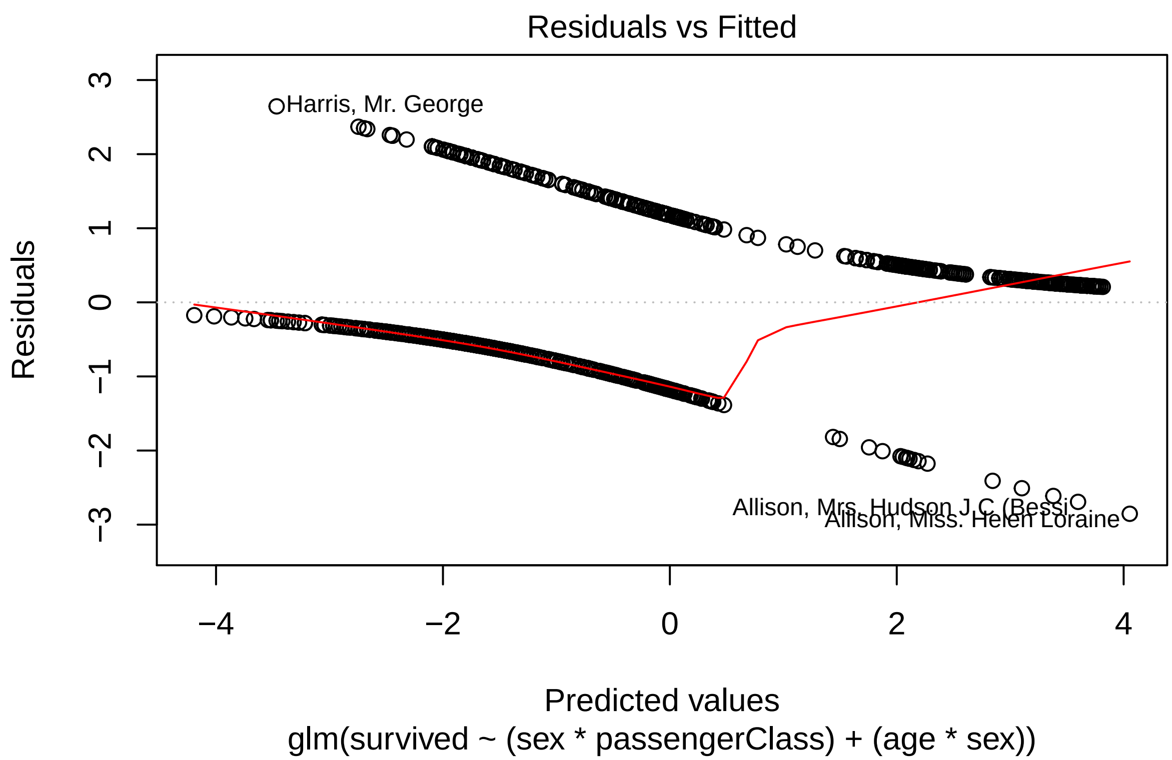
The final model is model3.

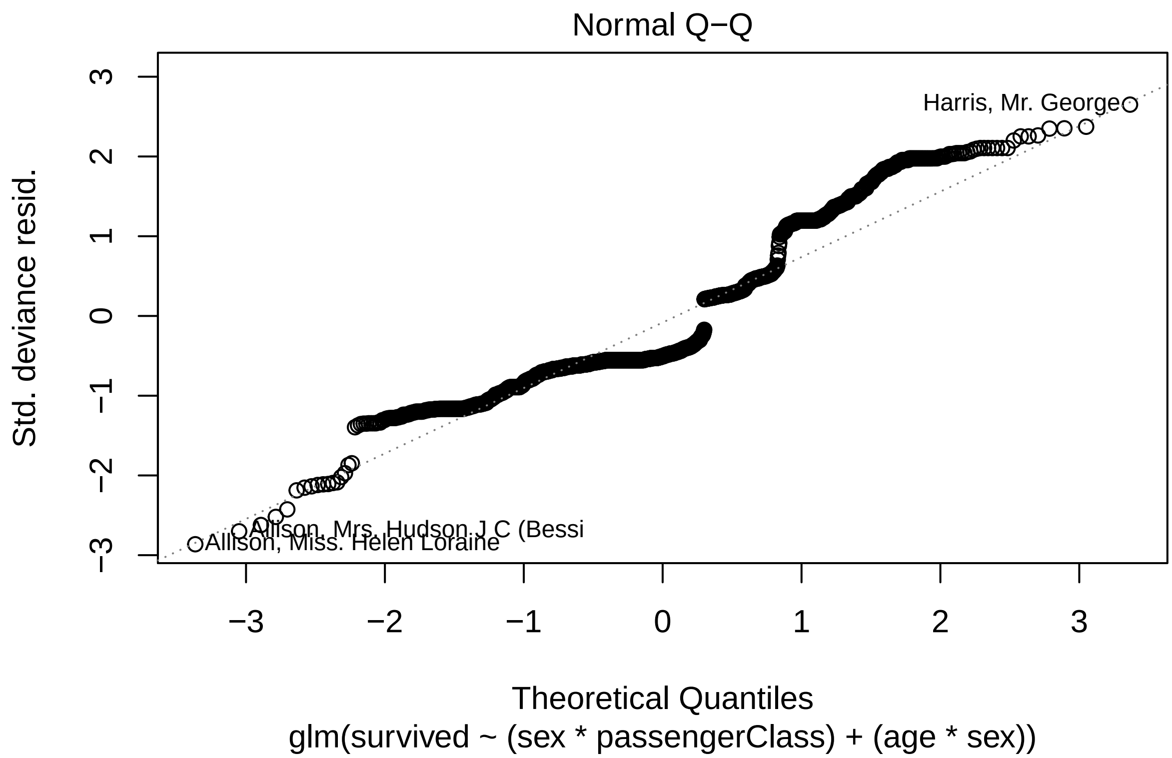
```r
Anova(model3, ty =3) # To see that everything is significant
```

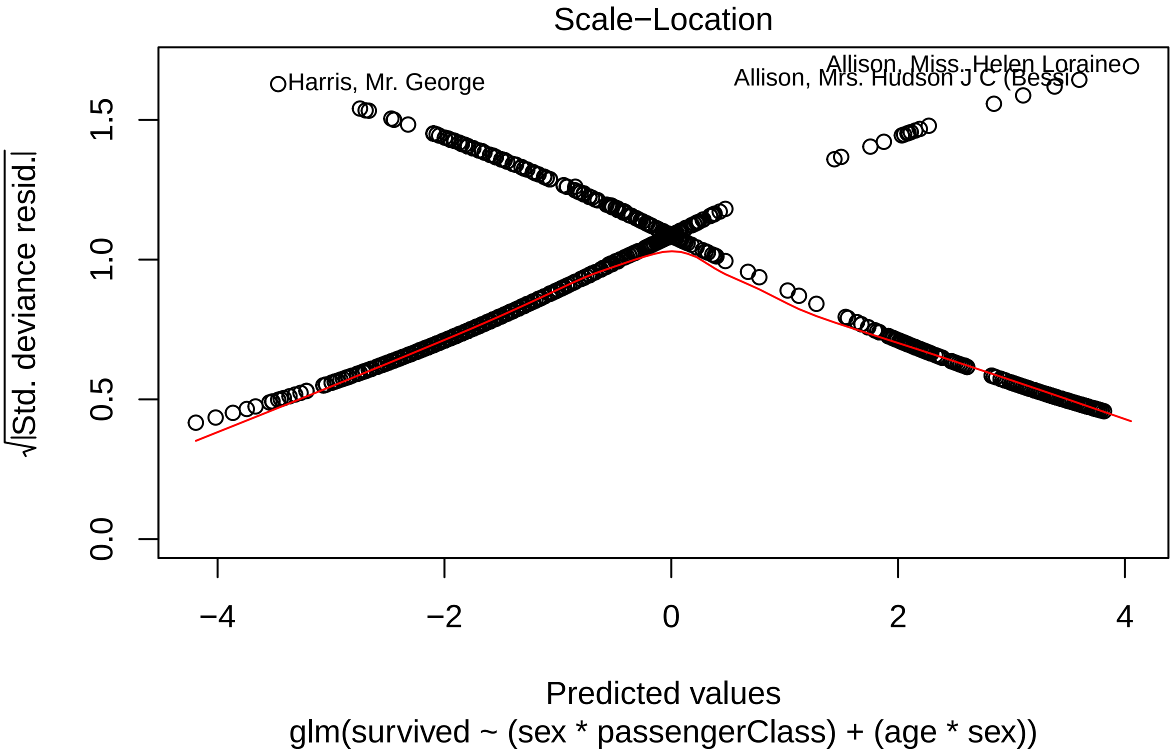
```
## Analysis of Deviance Table (Type III tests)
##
## Response: survived
##                      LR Chisq Df Pr(>Chisq)
## sex                    15.994  1  6.355e-05 ***
```

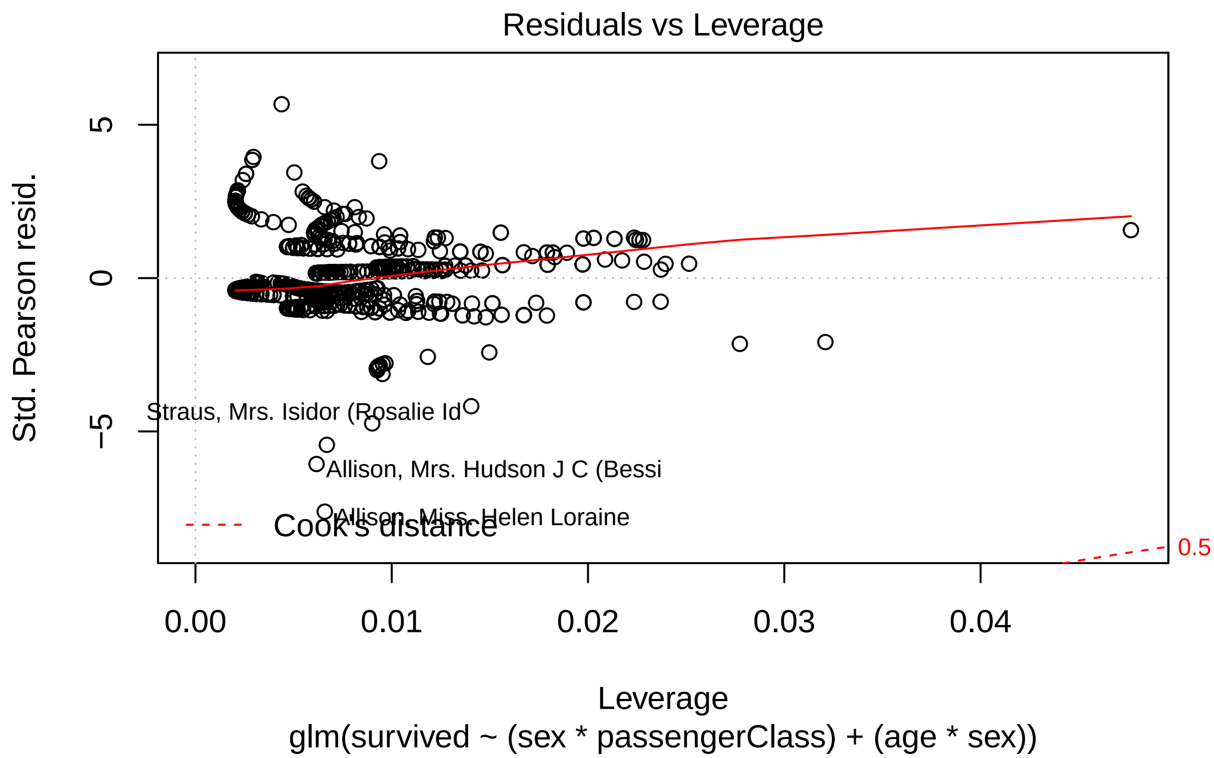
```
## passengerClass      119.767  2  < 2.2e-16 ***
## age                   3.196  1    0.07380 .
## sex:passengerClass   32.398  2  9.221e-08 ***
## sex:age               4.518  1    0.03355 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
plot(model3)
```




Normal Q−Q
Harris, Mr. George
Allison, Mrs. Hudson J C (Bessi
Allison, Miss. Helen Loraine
Std. deviance resid.
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3
Theoretical Quantiles
glm(survived ~ (sex * passengerClass) + (age * sex))


Scale−Location
Harris, Mr. George
Allison, Miss. Helen Loraine
Allison, Mrs. Hudson J C (Bessi
√|Std. deviance resid.|
1.5
1.0
0.5
0.0
−4
−2
0
2
4
Predicted values
glm(survived ~ (sex * passengerClass) + (age * sex))



TODO: There are clearly 2 lines. One for people that have survived and one for those that have not survived.

Interpret this first plot.

Plot the predicted values and plot the probabilities of model3 as function of the age variable, as our response variable.

Use diff. color for the (2 * 3 =) 6 diff profiles.

Take conclusions for this plot.

$predict_i = log(fracp_i 1 - p_i)$