Models for binary outcomes

SSSA - Applied Statistics - Chiara Seghieri and Costanza Tortù

2023-10-22

Load data

```
library(readxl)
library(foreign)
library(haven)
titanic <- as.data.frame(read_excel("~/Documents/Sant'Anna/Corso allievi/Data/TITANIC/TITANIC.xlsx"))
mytitanic<-titanic[,c("survived","pclass","age","Gender","cabin","sibsp","parch")]
```

Have a preliminary overview of data

look at the columns

```
ncol(titanic)
```

```
## [1] 15
```

```
head(mytitanic)
```

```
##   survived pclass age Gender cabin sibsp parch
## 1        0      3  42      0  <NA>     0     0
## 2        0      3  13      0  <NA>     0     2
## 3        0      3  16      0  <NA>     1     1
## 4        1      3  35      1  <NA>     1     1
## 5        1      3  16      1  <NA>     0     0
## 6        1      3  25      0 F G63     0     0
```

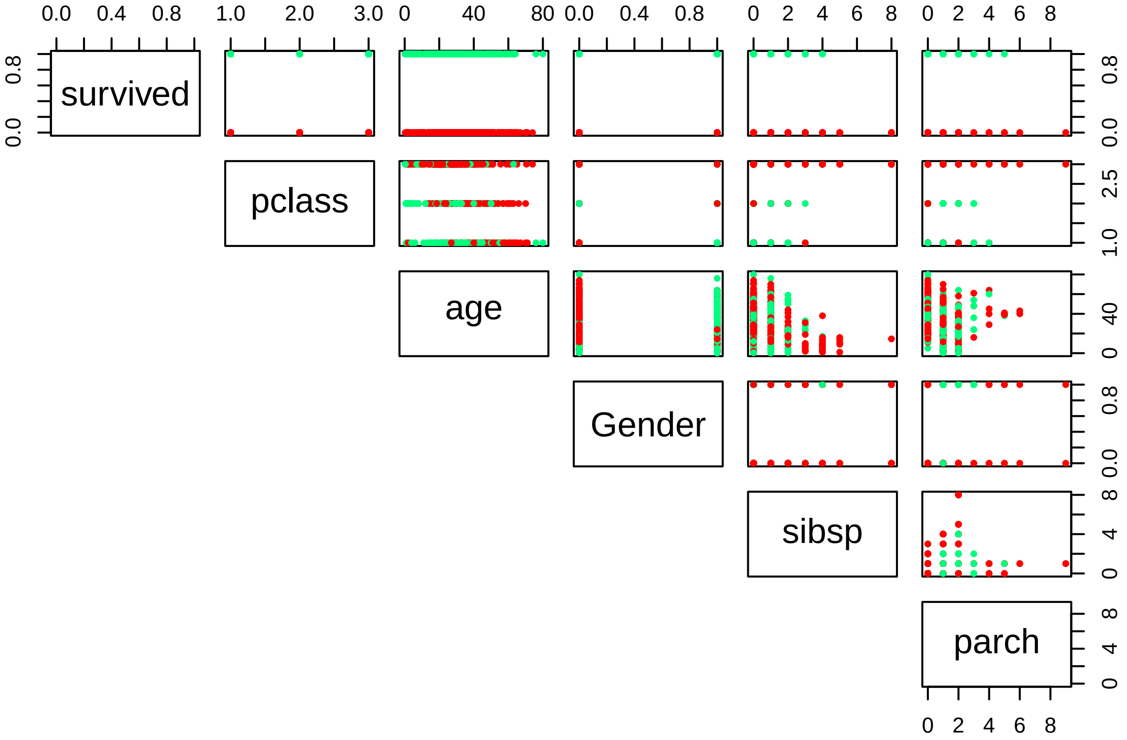
Are there missing entries?

```
summary(titanic)
```

```
##      pclass         survived       Residence        name          
##  Min.   :1.000   Min.   :0.000   Min.   :0.000   Length:1309       
##  1st Qu.:2.000   1st Qu.:0.000   1st Qu.:1.000   Class :character  
##  Median :3.000   Median :0.000   Median :2.000   Mode  :character  
##  Mean   :2.295   Mean   :0.382   Mean   :1.375                     
##  3rd Qu.:3.000   3rd Qu.:1.000   3rd Qu.:2.000                     
##  Max.   :3.000   Max.   :1.000   Max.   :2.000                     
##                                                                    
##       age             sibsp            parch          ticket          
##  Min.   : 0.1667   Min.   :0.0000   Min.   :0.000   Length:1309
##  1st Qu.:21.0000   1st Qu.:0.0000   1st Qu.:0.000   Class :character
##  Median :28.0000   Median :0.0000   Median :0.000   Mode  :character
##  Mean   :29.8811   Mean   :0.4989   Mean   :0.385
##  3rd Qu.:39.0000   3rd Qu.:1.0000   3rd Qu.:0.000
##  Max.   :80.0000   Max.   :8.0000   Max.   :9.000
##  NA's   :263
##       fare            cabin             embarked             boat
##  Min.   :  0.000   Length:1309        Length:1309        Length:1309
##  1st Qu.:  7.896   Class :character   Class :character   Class :character
##  Median : 14.454   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 33.295
##  3rd Qu.: 31.275
##  Max.   :512.329
##  NA's   :1
##       body        home.dest             Gender
##  Min.   :  1.0   Length:1309        Min.   :0.000
##  1st Qu.: 72.0   Class :character   1st Qu.:0.000
##  Median :155.0   Mode  :character   Median :0.000
##  Mean   :160.8                      Mean   :0.356
##  3rd Qu.:256.0                      3rd Qu.:1.000
##  Max.   :328.0                      Max.   :1.000
##  NA's   :1188
```

Inspect variables

```
my_cols <- c( "red","springgreen")
pairs(mytitanic[,c("survived","pclass","age","Gender","sibsp","parch")],
      pch = 19,  cex = 0.5,  col = my_cols[mytitanic$survived + 1],
      lower.panel=NULL)
```

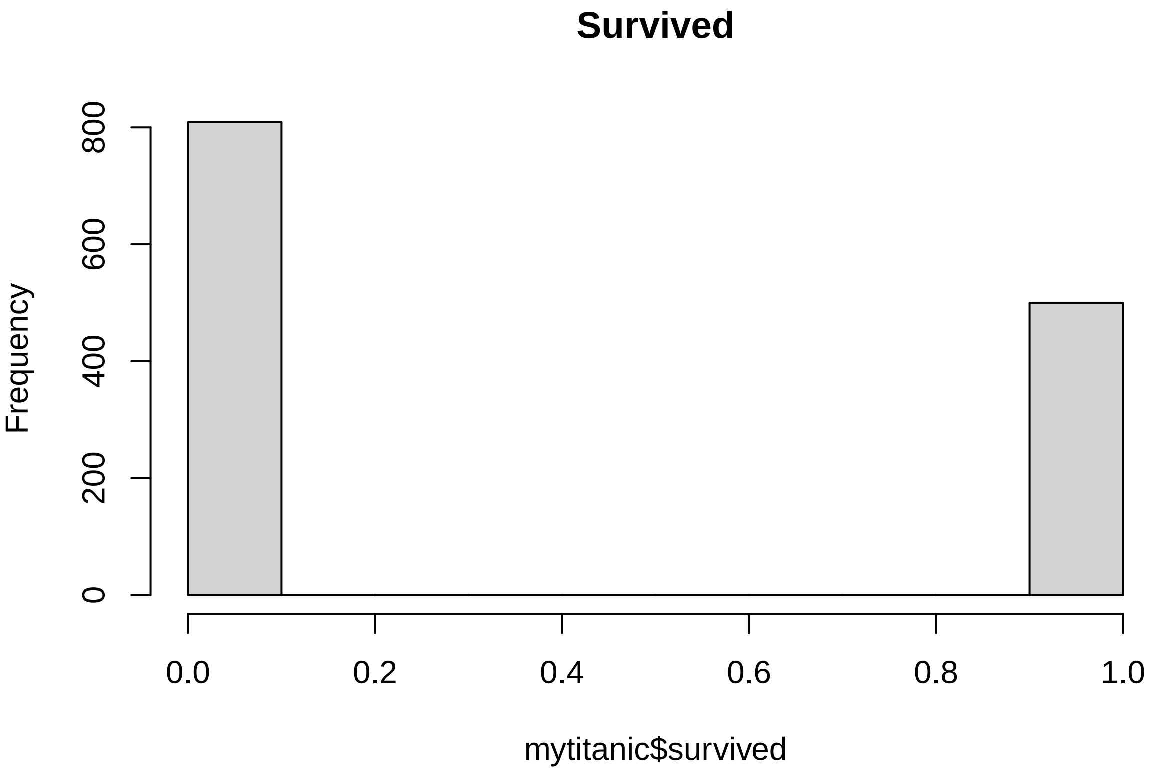


Inspect the outcome of interest

```
table(mytitanic$survived)
```

```
##
##   0   1
## 809 500
```

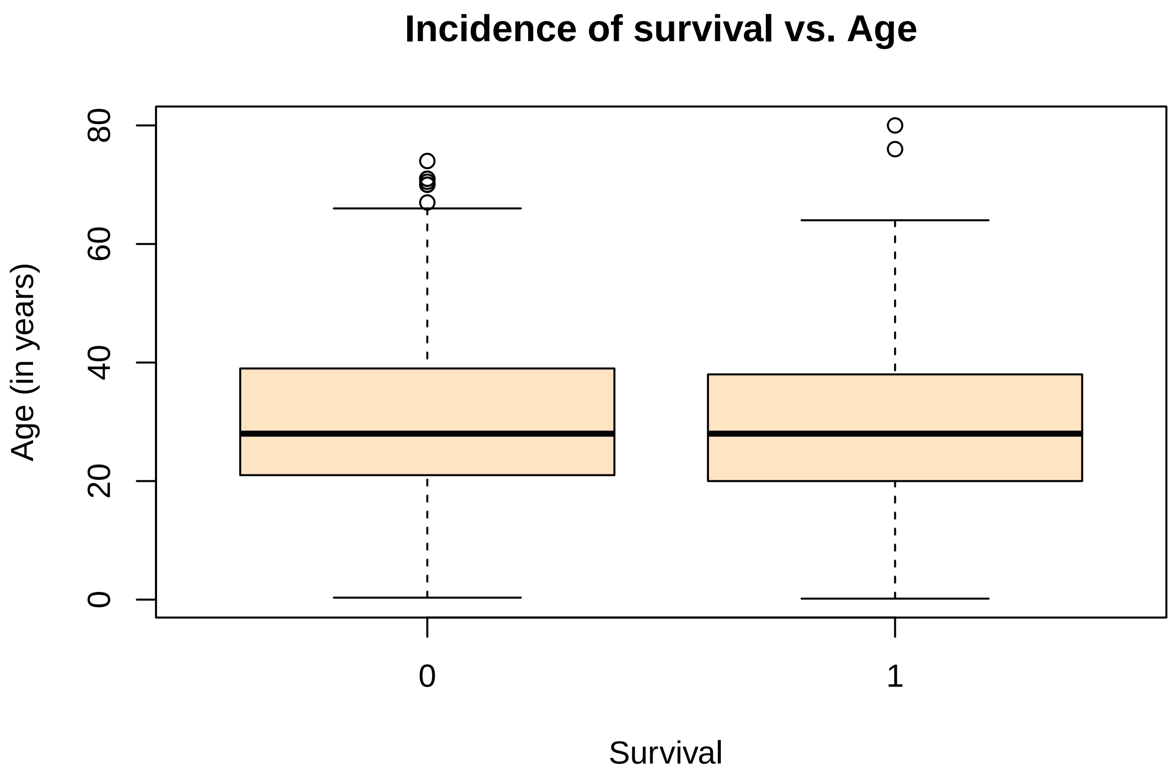
```
hist(mytitanic$survived, main = "Survived" )
```



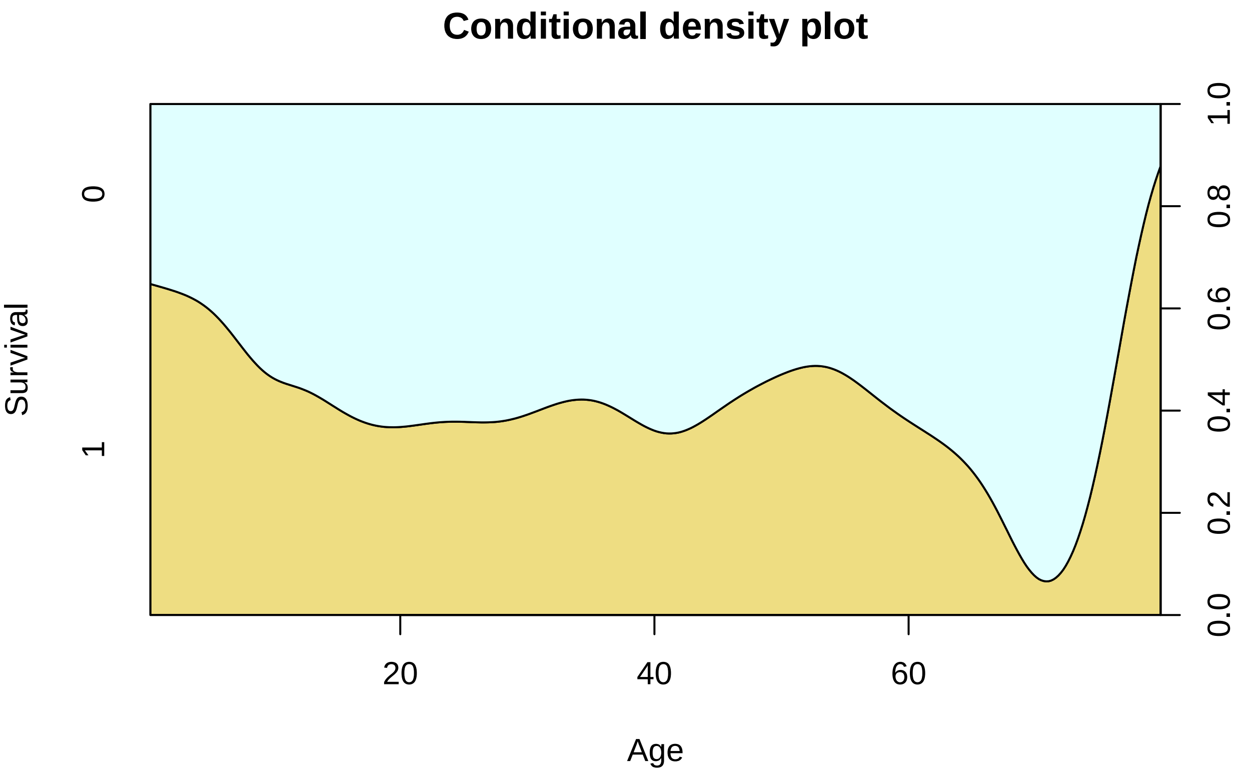
# Fit your models

First, we investigate the relationship between age and survival

```
boxplot(age ~ survived,data = mytitanic,
        col = "bisque",
        xlab = " Survival",
        ylab = "Age (in years)",
        main = "Incidence of survival vs. Age")
```



```
cdplot(as.factor(survived)~age,mytitanic,
       col=c("lightgoldenrod", "lightcyan"),
       ylab = "Survival",
       xlab ="Age", main = "Conditional density plot")
```



Compare the approaches for binary data with a naive linear approach

Naive linear model

```
simple_naive_lm <- lm(survived ~ age,
              data = mytitanic)
summary(simple_naive_lm)
```

```
##
## Call:
## lm(formula = survived ~ age, data = mytitanic)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4642 -0.4156 -0.3796  0.5806  0.6867
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.464814   0.034973  13.291   <2e-16 ***
## age         -0.001894   0.001054  -1.796   0.0727 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4912 on 1044 degrees of freedom
##   (263 observations deleted due to missingness)
## Multiple R-squared:  0.003082,   Adjusted R-squared:  0.002127
## F-statistic: 3.227 on 1 and 1044 DF,  p-value: 0.07271
```

```
simple_naive_lm_predicted <- predict(simple_naive_lm)
```

Simple Logit Model

```
simple_logit_fit <- glm(survived ~ age, family="binomial"(link="logit"),
                    data = mytitanic)
summary(simple_logit_fit)
```

```
##
## Call:
## glm(formula = survived ~ age, family = binomial(link = "logit"),
##     data = mytitanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1189  -1.0361  -0.9768   1.3187   1.5162
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.136531   0.144715  -0.943   0.3455
## age         -0.007899   0.004407  -1.792   0.0731 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

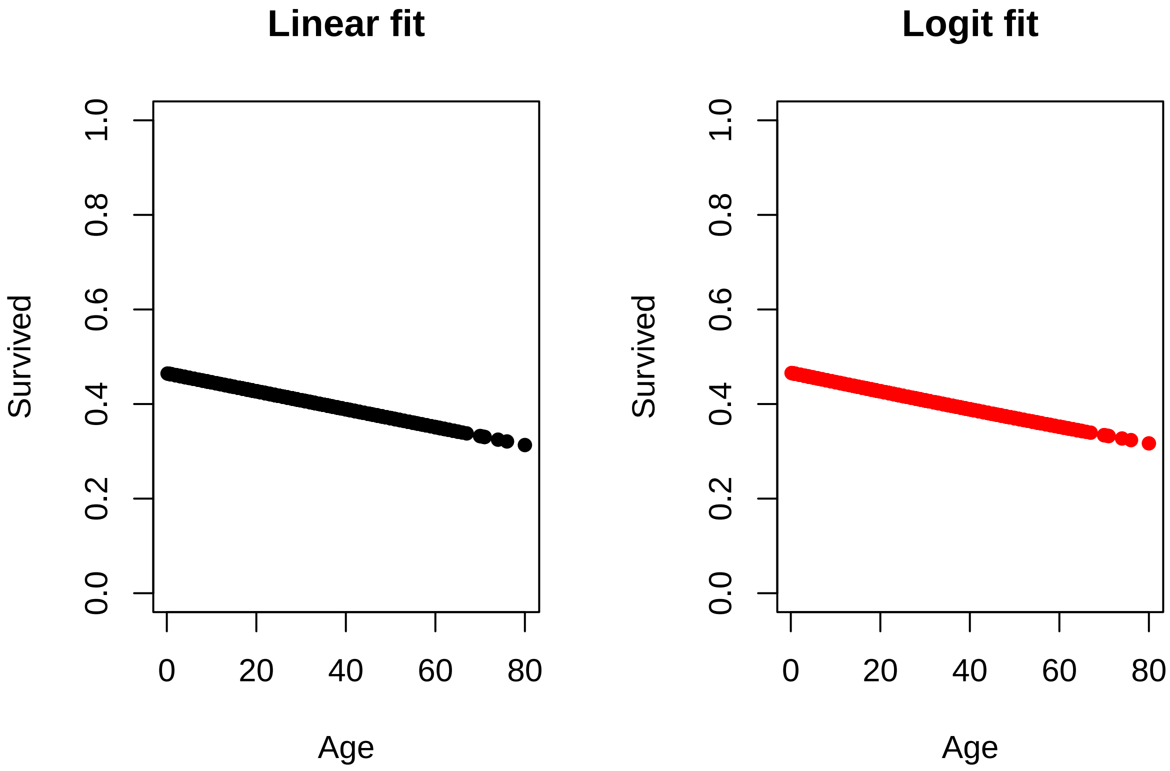
```
##
##     Null deviance: 1414.6  on 1045  degrees of freedom
## Residual deviance: 1411.4  on 1044  degrees of freedom
##   (263 observations deleted due to missingness)
## AIC: 1415.4
##
## Number of Fisher Scoring iterations: 4
```

```
simple_logit_fit_predicted <- predict(simple_logit_fit,type="response")
```

```
par(mfrow = c(1,2))

plot(mytitanic$age[which(!is.na(mytitanic$age))], main = "Linear fit",
     simple_naive_lm_predicted, xlab = "Age", ylab = "Survived",
     pch=16, col="black", ylim=c(0,1) )

plot(mytitanic$age[which(!is.na(mytitanic$age))], main ="Logit fit",
     simple_logit_fit_predicted, xlab = "Age", ylab = "Survived",
     pch=16, col="red", ylim=c(0,1) )
```



Here you do not see relevant differences but in some cases LPM can predict probabilities that are out of the range [0,1]!!!

Add a quadratic term for age

```
simple_logit_fit2 <- glm(survived ~ age + I(age^2), family="binomial"(link="logit"),
                        data = mytitanic)
summary(simple_logit_fit2)
```

```
##
```

```
## Call:
## glm(formula = survived ~ age + I(age^2), family = binomial(link = "logit"),
##     data = mytitanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2945  -1.0143  -0.9649   1.3422   1.4107
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.2844312  0.2331485   1.220  0.22248
## age         -0.0402649  0.0147200  -2.735  0.00623 **
## I(age^2)     0.0004955  0.0002144   2.311  0.02081 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1414.6  on 1045  degrees of freedom
## Residual deviance: 1406.1  on 1043  degrees of freedom
##   (263 observations deleted due to missingness)
## AIC: 1412.1
##
## Number of Fisher Scoring iterations: 4
```

```
simple_logit_fit_predicted2 <- predict(simple_logit_fit2,type="response")
```

Compare the two models

```
anova(simple_logit_fit,simple_logit_fit2,test='LR')
```

```
## Analysis of Deviance Table
##
## Model 1: survived ~ age
## Model 2: survived ~ age + I(age^2)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1044     1411.4
## 2      1043     1406.1  1   5.3253  0.02102 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Add covariates

Add other meaningful covariates

```
logit_fit <- glm(survived ~ age + I(age^2) + as.factor(pclass) +
                as.factor(Gender) +  sibsp + as.factor(parch),
                family = "binomial"(link="logit"),
                data = mytitanic)
summary(logit_fit)
```

```
##
## Call:
## glm(formula = survived ~ age + I(age^2) + as.factor(pclass) +
##     as.factor(Gender) + sibsp + as.factor(parch), family = binomial(link = "logit"),
##     data = mytitanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7435  -0.6565  -0.4160   0.6368   2.4479
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.433e+00  4.406e-01   3.253  0.00114 **
## age                 -6.113e-02  2.139e-02  -2.858  0.00426 **
## I(age^2)             4.018e-04  2.987e-04   1.345  0.17858
## as.factor(pclass)2 -1.312e+00  2.319e-01  -5.658 1.53e-08 ***
## as.factor(pclass)3 -2.227e+00  2.318e-01  -9.610  < 2e-16 ***
## as.factor(Gender)1  2.565e+00  1.749e-01  14.666  < 2e-16 ***
## sibsp               -4.548e-01  1.118e-01  -4.066 4.78e-05 ***
## as.factor(parch)1   6.950e-01  2.497e-01   2.783  0.00539 **
## as.factor(parch)2   3.173e-01  3.190e-01   0.995  0.31991
## as.factor(parch)3   6.641e-01  8.506e-01   0.781  0.43494
## as.factor(parch)4  -1.288e+00  1.262e+00  -1.020  0.30754
## as.factor(parch)5  -9.238e-01  1.161e+00  -0.796  0.42624
## as.factor(parch)6  -1.305e+01  5.271e+02  -0.025  0.98026
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1414.62  on 1045  degrees of freedom
## Residual deviance:  953.92  on 1033  degrees of freedom
##   (263 observations deleted due to missingness)
## AIC: 979.92
##
## Number of Fisher Scoring iterations: 13
```

Add interactions

```
logit_fit_winteract <- glm(survived ~ age + I(age^2) + as.factor(pclass) +
                as.factor(Gender) +  sibsp +  as.factor(parch)+ age*as.factor(Gender) +
                + as.factor(pclass)*as.factor(Gender),
                family = "binomial",
                data = mytitanic)
summary(logit_fit_winteract)
```

```
##
## Call:
## glm(formula = survived ~ age + I(age^2) + as.factor(pclass) +
##     as.factor(Gender) + sibsp + as.factor(parch) + age * as.factor(Gender) +
##     +as.factor(pclass) * as.factor(Gender), family = "binomial",
##     data = mytitanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9006  -0.6762  -0.4144   0.4483   2.4944
```

```
##
## Coefficients:
##                                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)                            1.944e+00  5.251e-01   3.702 0.000214
## age                                   -8.978e-02  2.403e-02  -3.736 0.000187
## I(age^2)                               6.106e-04  3.323e-04   1.838 0.066104
## as.factor(pclass)2                    -1.790e+00  3.280e-01  -5.456 4.86e-08
## as.factor(pclass)3                    -1.687e+00  2.782e-01  -6.064 1.33e-09
## as.factor(Gender)1                     2.488e+00  7.664e-01   3.246 0.001171
## sibsp                                 -4.758e-01  1.125e-01  -4.230 2.34e-05
## as.factor(parch)1                      6.821e-01  2.655e-01   2.569 0.010186
## as.factor(parch)2                      3.387e-01  3.386e-01   1.000 0.317153
## as.factor(parch)3                      3.554e-01  9.328e-01   0.381 0.703195
## as.factor(parch)4                     -1.805e+00  1.638e+00  -1.101 0.270681
## as.factor(parch)5                     -6.864e-01  1.169e+00  -0.587 0.557268
## as.factor(parch)6                     -1.300e+01  5.541e+02  -0.023 0.981284
## age:as.factor(Gender)1                 3.726e-02  1.519e-02   2.453 0.014149
## as.factor(pclass)2:as.factor(Gender)1  4.592e-01  6.629e-01   0.693 0.488464
## as.factor(pclass)3:as.factor(Gender)1 -1.893e+00  6.002e-01  -3.154 0.001609
##
## (Intercept)                           ***
## age                                   ***
## I(age^2)                              .
## as.factor(pclass)2                    ***
## as.factor(pclass)3                    ***
## as.factor(Gender)1                    **
## sibsp                                 ***
## as.factor(parch)1                     *
## as.factor(parch)2
## as.factor(parch)3
## as.factor(parch)4
## as.factor(parch)5
## as.factor(parch)6
## age:as.factor(Gender)1                *
## as.factor(pclass)2:as.factor(Gender)1
## as.factor(pclass)3:as.factor(Gender)1 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1414.62  on 1045  degrees of freedom
## Residual deviance:  897.48  on 1030  degrees of freedom
##   (263 observations deleted due to missingness)
## AIC: 929.48
##
## Number of Fisher Scoring iterations: 13
```

The probit link

```
probit_fit <- glm(survived ~ age + as.factor(pclass) +
                as.factor(Gender) +  sibsp + as.factor(parch),
                family = binomial(link = "probit"),
                data = mytitanic)
summary(probit_fit)
```

```
##
## Call:
## glm(formula = survived ~ age + as.factor(pclass) + as.factor(Gender) +
##     sibsp + as.factor(parch), family = binomial(link = "probit"),
##     data = mytitanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7735  -0.6795  -0.4262   0.6587   2.4453
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         0.585532   0.192799   3.037  0.00239 **
## age                -0.019199   0.003925  -4.891 1.00e-06 ***
## as.factor(pclass)2 -0.791050   0.132938  -5.951 2.67e-09 ***
## as.factor(pclass)3 -1.274137   0.129991  -9.802  < 2e-16 ***
## as.factor(Gender)1  1.516677   0.098153  15.452  < 2e-16 ***
## sibsp              -0.252546   0.062507  -4.040 5.34e-05 ***
## as.factor(parch)1   0.450404   0.138650   3.248  0.00116 **
## as.factor(parch)2   0.215156   0.180407   1.193  0.23302
## as.factor(parch)3   0.417637   0.501880   0.832  0.40533
## as.factor(parch)4  -0.801661   0.737423  -1.087  0.27699
## as.factor(parch)5  -0.635426   0.671874  -0.946  0.34428
## as.factor(parch)6  -4.344843  88.834732  -0.049  0.96099
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1414.62  on 1045  degrees of freedom
## Residual deviance:  957.08  on 1034  degrees of freedom
##   (263 observations deleted due to missingness)
## AIC: 981.08
##
## Number of Fisher Scoring iterations: 12
```

```
deviance(probit_fit)
```

```
## [1] 957.081
```

```
deviance(logit_fit)
```

```
## [1] 953.9152
```

```
d <- deviance(probit_fit) - deviance(logit_fit)
d
```

```
## [1] 3.165827
```