Estimating a logistic regression model in R

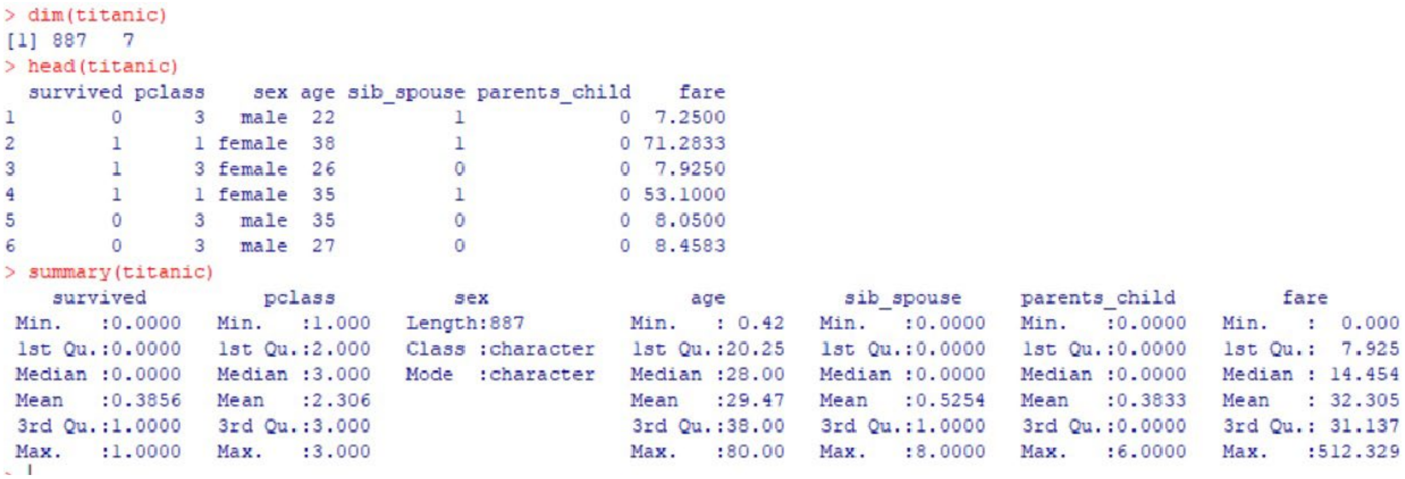
Example: Who survived the Titanic?

There were an estimated 2,224 passengers and crew aboard the Titanic, and more than 1,500 died, making it one of the deadliest commercial peacetime maritime disasters in modern history. Can we predict who survived using passenger records?

The data set has n=887 passenger records with 7 variables, 39% are survivors

```
dim(titanic)            #gives the dimension of the data set (rows, columns)
head(titanic)           #prints the first few rows of the data set
summary(titanic)        #summarizes each variable in the data set
```

Which yields the following:



```
> dim(titanic)
[1] 887   7
> head(titanic)
  survived pclass    sex age sib_spouse parents_child    fare
1        0      3   male  22          1             0  7.2500
2        1      1 female  38          1             0 71.2833
3        1      3 female  26          0             0  7.9250
4        1      1 female  35          1             0 53.1000
5        0      3   male  35          0             0  8.0500
6        0      3   male  27          0             0  8.4583
> summary(titanic)
    survived          pclass          sex                 age          sib_spouse     parents_child        fare
 Min.   :0.0000   Min.   :1.000   Length:887         Min.   : 0.42   Min.   :0.0000   Min.   :0.0000   Min.   :  0.000
 1st Qu.:0.0000   1st Qu.:2.000   Class :character   1st Qu.:20.25   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:  7.925
 Median :0.0000   Median :3.000   Mode  :character   Median :28.00   Median :0.0000   Median :0.0000   Median : 14.454
 Mean   :0.3856   Mean   :2.306                      Mean   :29.47   Mean   :0.5254   Mean   :0.3833   Mean   : 32.305
 3rd Qu.:1.0000   3rd Qu.:3.000                      3rd Qu.:38.00   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.: 31.137
 Max.   :1.0000   Max.   :3.000                      Max.   :80.00   Max.   :8.0000   Max.   :6.0000   Max.   :512.329
```

We can see that there are a few possible predictors of surviving the Titanic:

pclass – which represents passenger class and can take the values 1,2,or 3. 3rd class is the lowest class, while 1st class is the most luxurious class.

sex – biological sex

sib_spouse – provides the number of siblings + spouse travelling with the passenger (8 was the max in this data set)

parents_child – provides the number of parents + children traveling with the passenger (6 was the max)

fare – the price of the ticket (0 was the minimum, 32.305 was the average fare, and 512.329 was the maximum). Of course, fare will be correlated with passenger class. We can see this relationship by looking at the average and maximum fare for each passenger class:

```
library(data.table)
setDT(titanic)[ , .(mean_fare =
mean(fare),max_fare=max(fare),min_fare=min(fare)), by = pclass]
```

Which yields:

```
  pclass      mean_fare    max_fare    min_fare

1:    3       13.70771     69.55000      0

2:    1       84.15469     512.3292      0

3:    2       20.66218     73.5000       0
```

As expected, passengers in 1st class paid an average of 84.2 pounds for their ticket, while 3rd class passengers paid 13.7 pounds on average. It is not a perfect correlation however – as you can see some passengers in 3rd class actually paid more for their ticket (max fare=69.6) than those in 1st class (min fare=0).

Now suppose we want to estimate a simple logistic regression model consisting of three predictors of interest – passenger class, biological sex, and age. We assume that these three predictors will explain most of the variability in who survived the Titanic sinking.

To fit a logistic regression model:

```
model <- glm(survived~pclass + sex + age, family="binomial",
data=titanic)

summary(model)
```

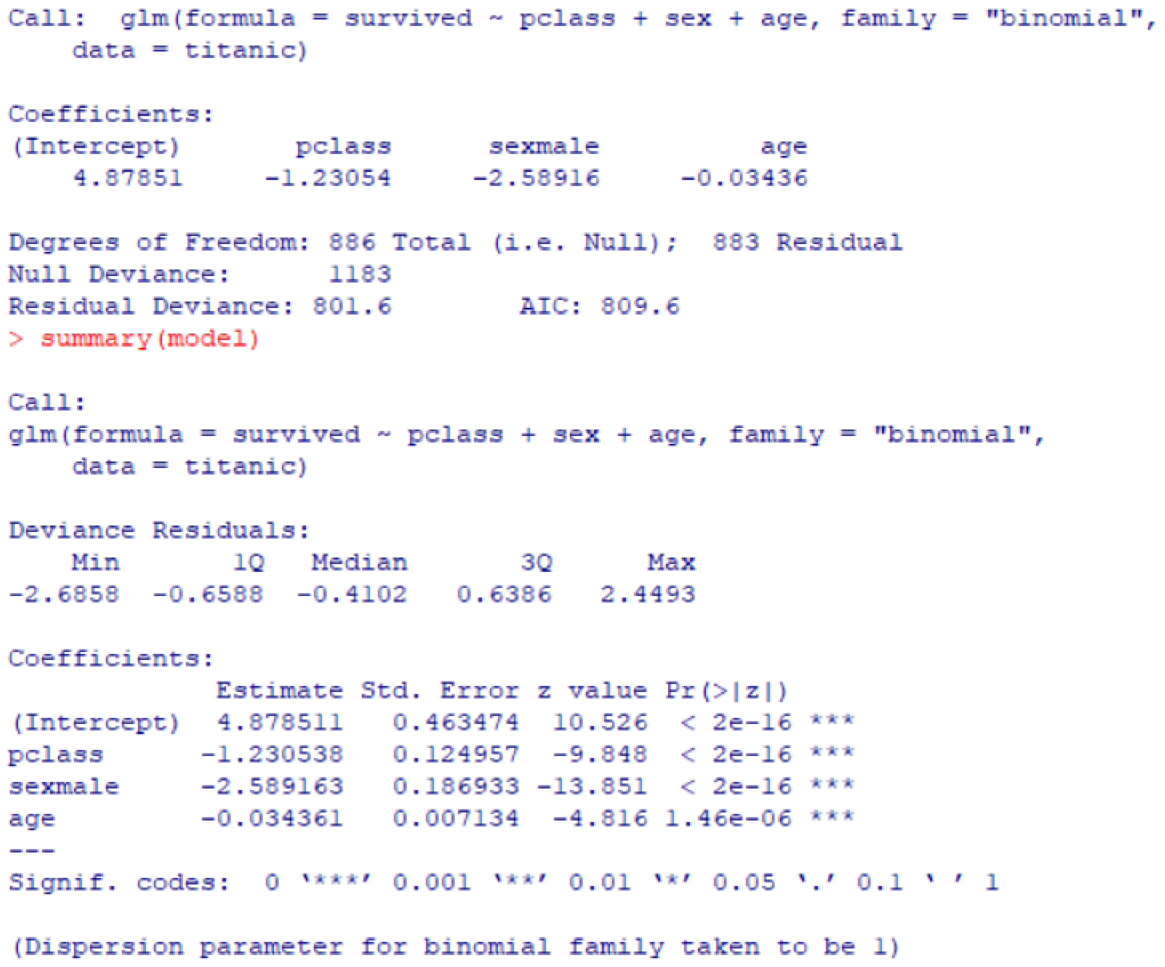
Notes:

1) The syntax in glm is always:

dependent variable ~ independent variable1 + *independent variable 2* +……

2) *family* = *binomial* tells R to expect a binary outcome (dependent variable is survived, which is 0/1) and to perform logistic regression (as opposed to another type of regression, like linear or Poisson)

The R output is below:



```
Call:  glm(formula = survived ~ pclass + sex + age, family = "binomial",
    data = titanic)

Coefficients:
(Intercept)       pclass      sexmale          age
    4.87851     -1.23054     -2.58916     -0.03436

Degrees of Freedom: 886 Total (i.e. Null);  883 Residual
Null Deviance:      1183
Residual Deviance: 801.6        AIC: 809.6
> summary(model)

Call:
glm(formula = survived ~ pclass + sex + age, family = "binomial",
    data = titanic)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6858  -0.6588  -0.4102   0.6386   2.4493

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.878511   0.463474  10.526  < 2e-16 ***
pclass      -1.230538   0.124957  -9.848  < 2e-16 ***
sexmale     -2.589163   0.186933 -13.851  < 2e-16 ***
age         -0.034361   0.007134  -4.816 1.46e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

Some things to notice in this output:

1. The estimated coefficients are given in the log odds scale. If we want odds ratios, we need to manipulate these with a few more lines of code (see below).
2. Also, the 95% confidence intervals are not provided with the summary command so we will extract those below too.
3. The p-values for each estimated coefficient are statistically significant (all with p-values < 0.00001)
4. The deviance residuals are computed for each observation and the median residual was equal to -0.4102. More generally, residuals are the differences between what we observe and what our model predicts. We generally would like the median deviance residual to be close to 0. In addition, we would like to see the minimum and maximum values be less than about 3 in absolute value.

Obtaining estimated odds ratios:

```
exp(coef(model))
```

```
 (Intercept)          pclass       sexmale      age
131.43485141          0.29213545   0.07508286   0.96622221
```

Obtaining 95% confidence intervals for the estimated odds ratios:

```
exp(confint(model))
```

```
Waiting for profiling to be done...
                2.5 %         97.5 %
(Intercept)    54.25666660   334.4556204
pclass         0.22741168    0.3713753
sexmale        0.05164634    0.1075586
age            0.95257335    0.9796254
```

Now that we have the estimated odds ratios along with their 95% confidence intervals, we can see that:

1. Male passengers had a much lower odds of surviving the Titanic (OR=0.075, 95% CI: (0.052,0.108)). The odds of survival is about 8% that of the odds of survival for female passengers
2. As passenger class increases by one unit (1st to 2nd or 2nd to 3rd), the odds of surviving decrease (OR=0.29, 95% CI: (0.23,0.37)). Someone in 3rd class has about one third the odds of surviving compared to someone in 2nd class.
3. As age increases, the odds of surviving also decreases. So it appears that the idea of **women and children first** holds true in the Titanic disaster.

How well does this model actually fit?

To get a sense of this, we can compute something called the McFadden's pseudo R^2 value using the pscl package. This value will be between 0 and 1 and higher values means the model is doing a good job in fitting the data.

```
library(pscl)
pscl::pR2(model)["McFadden"]
```

The resulting output is:

```
fitting null model for pseudo-r2
 McFadden
0.3222581
```

Note: A value of 0.20-0.40 is considered a good fit, while values > 0.40 are considered a very strong fit. As our value is 0.32, our simple model is actually fitting the data pretty well.