WiMa-Praktikum II – Stochastic Summer term 2016
Arthur Allignol 9.6.2016

Exercise Sheet 5 — Generalized Linear Model

Problem 1. Titanic Passenger Survival
The data set `titanic.csv` contains information on the fate of 891 passengers of the Titanic. Additionally, we have for each passenger the following information

PassengerId Passenger ID

Pclass Passenger Class

Name Passenger name

Sex Gender

Age Age

SibSp Number of Siblings/Spouses Aboard

Parch Number of Parents/Children Aboard

Ticket Ticket number

Fare Passenger fare

Cabin Cabin number

Embarked Port of embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

The variable `Survived` contains the passenger survival indicator (1 for survivors). The aim of the exercise is to predict the passengers survival.

(a) `Age` has 177 missing values. Perform a crude imputation by replacing the missing ages by the median age computed for each `Pclass`. Motivate this approach.

(b) Most families were given one ticket for all family members, whose price is recorded in variable `Fare`. Divide the fare by the number of individuals sharing the same ticket number.

(c) Find a good model that predicts survival (Hint: You can compare nested models with `anova(model1, model2, test = “Chisq”)`. Otherwise you can look at the AIC in the `summary` output. The AIC is a function of the log-likelihood and the number of parameters, thus provides a trade-off between complexity of the model and fit. The smaller the better. And don’t forget simple graphical analyses.)

(d) Based on your chosen model, predict the fate of Jack Dawson (Leonardo DiCaprio) and Rose DeWitt Bukater (Kate Winslet)