Dear all, this is an open book exam. Please submit your RMD output file in PDF or Word Format, to the class Brightspace site. Academic dishonesty will result in a grade of F for the class.

Logistic Regression with the Titanic Data – Classification Task

The Titanic2.csv data we will use for our quiz has 891 passengers and 12 variables:

- PassengerId: Passenger ID: 1- 891
- Survived: A binary variable indicating whether the passenger survived or not (0 = No; 1 = Yes); this is our response variable
- Pclass: Passenger class (1 = 1st; 2 = 2nd; 3 = 3rd)
- Name: A field rich in information as it contains title and family names
- Sex: male/female
- Age: Age, asignificant portion of values are missing
- SibSp: Number of siblings/spouses aboard
- Parch: Number of parents/children aboard
- Ticket: Ticket number.
- Fare: Passenger fare (British Pound).
- Cabin: Cabin number
- Embarked: Port of embarkation (C = Cherbourg; Q = Queenstown; S

= Southampton)

First, one must clean the data and decide which variables to exclude from our analysis. My recommendation is that we exclude PassengerID, Name, Ticket, and Cabin in the ensuing analysis. Next, please note that Age has many missing values – my suggestion is to delete those with missing values. Now after the data cleaning step, your task is to split the data randomly into training (80%) and testing (20%), first build the model using the training data, and then use that model to predict whether each passenger in the testing data survived or not. In the last part, we gave you one additional passenger (# 892) for prediction.

1. For the entire dataset, please perform the data cleaning as instructed before; namely, exclude the variables PassengerID, Name, Ticket, and Cabin and delete missing values in the variable Age. Please report how many passengers are left after this step.
2. Please use the "str()" function to examine the variable type in our data set. Please note that R will recognize a character valued function as categorical variable (namely a Factor variable) automatically and therefore you do not need

to transform this type of variables. However, you do need to recode the numerical valued categorical variables using the "as.factor()" function so that R can recognize these as categorical variables. Now, please recode the "Survived" and the "Pclass" variables as factors.

3. Now please use the random seed 123 to divide the cleaned data into 80% training and 20% testing.
4. Then you shall fit a logistic regression model with all the other 7 predictors using the training data.
5. Please use this fitted model based on the training data to predict the response variable ''Survived" (whether the subject survived or not) for the testing data. Please generate the confusion matrix, and report:
    a. The overall accuracy;
    b. The sensitivity (that is, the probability a subject is predicted to have survived given that he/she had survived);
    c. The specificity (that is, the probability a subject is predicted to have not survived given that he/she had not survived).
6. Now we have recovered the record of additional passengers as follows. Please predict whether these passengers have survived or not.

PassengerId	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabine	Embarked
892	3	male	24	1	0	123456	8.42		Q
893	1	female	68	0	0	123457	24.34	D41	C
894	2	male	41	1	2	123458	41.93		S