

# Department of Statistics

## COURSE STATS 330

Assignment 1, 2007

Instructions: Hand in your completed assignment to the Student Resource Centre by 4pm on Thursday 2<sup>nd</sup> August.

The data set in the file **titanic.txt** (available on the course web page) contains some data on 633 passengers on the liner Titanic, which sank in the North Atlantic on 15<sup>th</sup> April 1912 after striking an iceberg.

The data set has 5 variables and 633 cases. The variables are

**age.group:** The age group of the passenger (0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79), treated as a factor;

**age:** The age of the passenger, treated as a continuous variable;

**survived:** 0 = died, 1 = survived. This is a numeric variable.

**pclass:** The passenger class (1st, 2nd, 3rd), treated as a factor;

**sex:** The gender (female, male) of the passenger.

### Questions

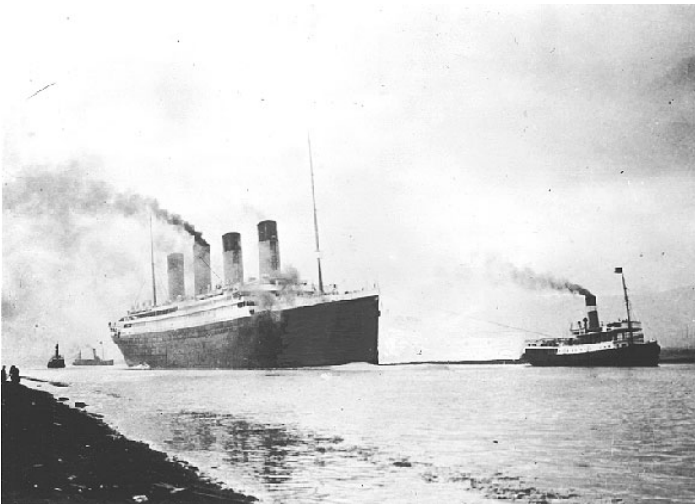
1. Load the data into R, and make a data frame **titanic.df** to contain the data. Check for any typographical errors. [5 marks]
2. Make an additional variable by turning the numeric variable **survived** into a factor **survival** with levels "survived" and "died". Add it to the data frame. [5 marks]
3. What is the relationship between survival and age? Does it depend on class and gender? Draw a suitable trellis plot to answer this question. Don't try and fit any models. [10 marks, 5 for the plot and 5 for the discussion]
4. For each combination of age group, class and gender, calculate the fraction of passengers that survived. Present your results in a table. [10 marks]

**PTO**

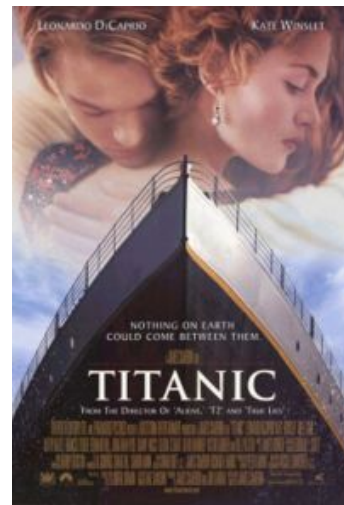
5. How does the fraction surviving depend on age group, gender and class? Draw another Trellis plot to explore this. [10 marks, 5 for the plot and 5 for the discussion]

## R hints

1. Use **read.table** to create the data frame.
2. Use **factor** to change the variable surviving into a factor. The argument **labels** can be used to set the labels to "survived" and "died".
3. Don't forget to label the axes. Use **xlab** and **ylab**. Skim the paper "A Tour of Trellis Graphics" for helpful hints.
4. Use the R function **table** to draw tables. Note some table entries will be NA (no passengers with that agegroup/sex/class combination).
5. You will have to convert the table into a data frame with 48 rows to draw the trellis plot. You can do this in R using the **data.frame** function, and **expand.grid** to create the factors, but this is a bit tricky. You might find converting the table into a vector (use **as.vector**) and pasting the result into Excel easier, or even just retype the numbers. You can fill in the factor levels using the Excel tools, then read the data back into R.



The real version



The movie version