Department of Statistics

COURSE STATS 330

Model answers for Assignment 1, 2007.

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 15th 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+), 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]*

There are several ways to do this. You can download the file titanic.txt onto your computer, and set the R directory to point to the folder containing the data. You set the R directory by pulling down the File menu in R, choosing “Change dir…”, and navigating to the correct folder. Having set the directory, type

```
titanic.df = read.table(“titanic.txt”, header=T)
```

Another way that is more convenient is to load the data directly from the web site. (You have to be connected to the internet to do this.) Type

```
titanic.df =
read.table("http://www.stat.auckland.ac.nz/~lee/330/datasets.dir/titanic.
txt", header=T)
```

You can cut and paste the URL for the data directly from the browser. To check for typographical errors, we can just inspect the values of the variables to see if there are any typos. You can proofread the file, but a simpler way is just to inspect the unique (i.e. distinct) values of the variables. Type

```
> unique(titanic.df$age.group)
[1] 20-29  0-9    30-39  40-49  60-69  50-59  70-79  10-19  100-19
Levels: 0-9 10-19 100-19 20-29 30-39 40-49 50-59 60-69 70-79
```

Here we see that there is a typo : 100-19 is not a valid age group. It probably should be 10-19, so we will correct it to that. Which case is the offending one?

By typing

```
> titanic.df[titanic.df$age.group=="100-19",]
    pclass survived age  sex age.group
633    3rd        0  19 male    100-19
```

we see that it is case number 633. Correct the original file and read the data in again Alternatively, we can correct it by

```
titanic.df[633,5]="10-19"
```

We also have to readjust the factor to eliminate the incorrect level:

```
titanic.df$age.group = factor(titanic.df$age.group)
```

since **age.group** is the 5th variable in the data frame. The other data can be checked similarly, but are OK.

2. *Make an additional variable by turning the numeric variable **survived** into a factor with levels "survived" and "died". [5 marks]*

Let's call the new variable **survival**. Make it by typing

```
survival = factor(titanic.df$survived, labels=c("Died", Survived"))
```

Add the new variable to the data frame, calling the result titanic2.df

```
titanic2.df = data.frame(survival, titanic.df)
```

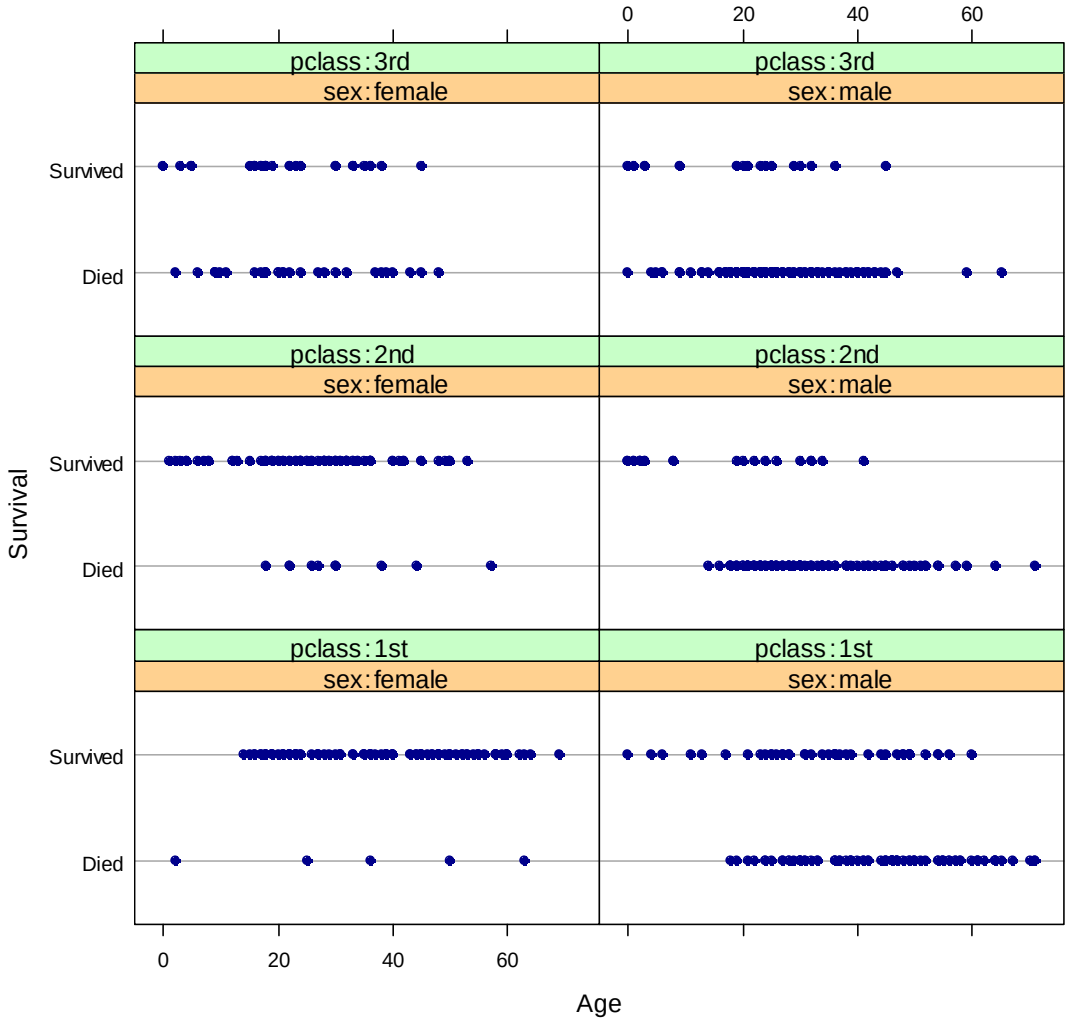
3. *What is the relationship between survival and age? Does it depend on class and gender? Draw a suitable trellis plot to answer this question. Don't try and fit any models. [10 marks, 5 for the plot and 5 for the discussion]*

We will draw a trellis plot of survival versus age, using gender and class as conditioning factors.

```
library(lattice)  # need to load the lattice library

trellis.par.set(background = list(col = "white"),
dot.symbol=list(col="darkblue"))

# trellis.par.set is the method for changing the background
# colours and plot symbol colours (this is optional, but these colours
are easier to see when printed)
```



Interpretation: For the third class passengers, the ages seem similar for those who died and those who survived. Note that relatively few third class passengers survived. For the second and first class passengers, the survivors tended to be younger than those who died, with the exception of female first class passengers. (There are too few of these who died for any pattern to emerge.)

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]*

The R function table is useful for this.

```
my.table =
table(titanic2.df$age.group,titanic2.df$pclass,titanic2.df$sex,
titanic2.df$survival)

> my.table
```

```
, ,  = female,  = Died

      1st 2nd 3rd
  0-9     1   0   5
  10-19   0   1   7
  20-29   1   4   8
  30-39   1   3   5
  40-49   0   1   4
  50-59   1   1   0
  60-69   1   0   0
  70-79   0   0   0

, ,  = male,  = Died

      1st 2nd 3rd
  0-9     0   0   7
  10-19   2  10  21
  20-29  10  43  53
  30-39  18  30  23
  40-49  22  13  14
  50-59  16   8   1
  60-69  11   1   1
  70-79   3   1   0

, ,  = female,  = Survived

      1st 2nd 3rd
  0-9     0   9   4
  10-19  13  11  11
  20-29  20  23   4
  30-39  19  19   8
  40-49  19   9   1
  50-59  18   4   0
  60-69   7   0   0
  70-79   0   0   0

, ,  = male,  = Survived

      1st 2nd 3rd
  0-9     3  11   6
  10-19   3   1   2
  20-29  10   4   6
  30-39  12   4   3
  40-49  10   1   1
  50-59   4   0   0
  60-69   1   0   0
  70-79   0   0   0
```

The object **my.table** is an array – this is like a matrix but in this case has 4 dimensions. We can make separate tables of just the survivors and just those

who survived by subsetting:

```
> survivors=my.table[,,,2]  # survivors are second level
> died = my.table[,,,1]  # died are first level
> fraction = survivors/(died + survivors)
> round(fraction,3) # rounds to 3 decimal places

, ,  = female

         1st   2nd   3rd
  0-9   0.000 1.000 0.444
  10-19 1.000 0.917 0.611
  20-29 0.952 0.852 0.333
  30-39 0.950 0.864 0.615
  40-49 1.000 0.900 0.200
  50-59 0.947 0.800   NaN
  60-69 0.875   NaN   NaN
  70-79   NaN   NaN   NaN

, ,  = male

         1st   2nd   3rd
  0-9   1.000 1.000 0.462
  10-19 0.600 0.091 0.087
  20-29 0.500 0.085 0.102
  30-39 0.400 0.118 0.115
  40-49 0.312 0.071 0.067
  50-59 0.200 0.000 0.000
  60-69 0.083 0.000 0.000
  70-79 0.000 0.000   NaN
```

An alternative way to make the table is to use the fact that, for binary (0/1) data like the variable survived, the proportion of ones is just the mean. We can calculate the mean (ie the proportion surviving) for each age group/class./sex combination by using the R function tapply:

```
> survival.frac = tapply(titanic2.df$survived,
  list(titanic2.df$age.group,titanic2.df$pclass,titanic2.df$sex), mean)
> round(survival.frac,3)



, , female

        1st   2nd   3rd
0-9   0.000 1.000 0.444
10-19 1.000 0.917 0.611
20-29 0.952 0.852 0.333
30-39 0.950 0.864 0.615
40-49 1.000 0.900 0.200
50-59 0.947 0.800    NA
60-69 0.875    NA    NA
70-79    NA    NA    NA
```

```
, , male

        1st   2nd   3rd
0-9   1.000 1.000 0.462
10-19 0.600 0.091 0.087
20-29 0.500 0.085 0.102
30-39 0.400 0.118 0.115
40-49 0.312 0.071 0.067
50-59 0.200 0.000 0.000
60-69 0.083 0.000 0.000
70-79 0.000 0.000    NA
```

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]*

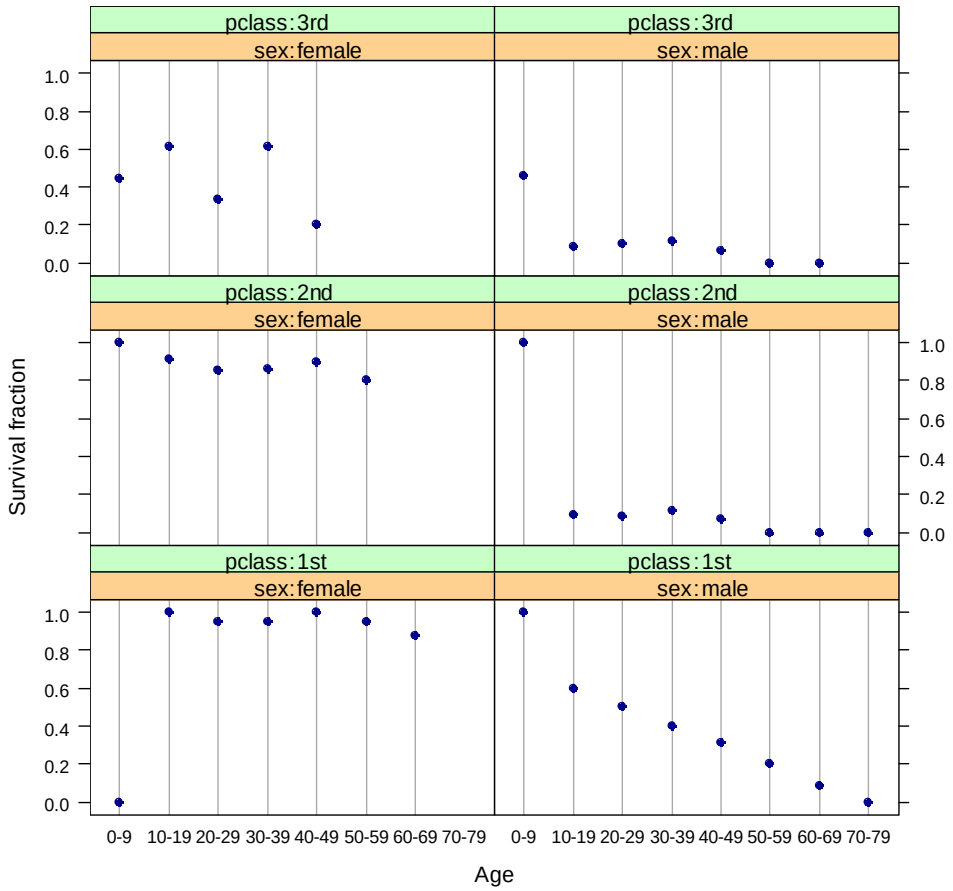
First, we need to convert the table sentries into a data frame, with extra variables indicating the factor levels. The R function expand.grid does this job nicely:

```
survival.frac.df = data.frame(frac = as.vector(survival.frac), expand.grid(
age.group=c("0-9","10-19","20-29","30-39","40-49","50-59","60-69","70-79"),
pclass=c("1st","2nd","3rd"),
sex=c("female", "male")))
```

```
> survival.frac.df
    frac age.group pclass    sex
1  0.000       0-9    1st   male
2  1.000     10-19    1st   male
3  0.952     20-29    1st   male
4  0.950     30-39    1st   male
5  1.000     40-49    1st   male
6  0.947     50-59    1st   male
7  0.875     60-69    1st   male
8     NA     70-79    1st   male
9  1.000       0-9    2nd   male
10 0.917     10-19    2nd   male
11 0.852     20-29    2nd   male
12 0.864     30-39    2nd   male
13 0.900     40-49    2nd   male
14 0.800     50-59    2nd   male
15    NA     60-69    2nd   male
16    NA     70-79    2nd   male
17 0.444       0-9    3rd   male
…….   48 lines in all
```

Then we draw the dot plot

```
dotplot(frac~age.group|sex*pclass,
data=survival.frac.df,
xlab="Age",
ylab="Survival fraction",
strip=function(...)strip.default(..., strip.names=T))
```



Discussion: It is clear from the plot that survival was much higher for females than males for all classes, although the higher the class, the higher the survival. Survival was more likely for younger persons. This trend is particularly pronounced for the first class males. Age did not have such a strong effect on survival for the other categories.

Sounds like the movie got it right!