P09 – Description of the Titanic data

1.1 Data source

The Titanic passenger data is freely available on the internet in various versions; we use the one from here (don’t download it yet – you’ll get a pre-processed version):

- Data: http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/titanic3.csv
- Description: http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/titanic3info.txt

Our data has been processed by the following script in order to make it easier to work with using Python (version 2.x!) and Excel (it’s here just for reference reasons):

```
import re

print "Convertig file..."

with open('titanic3.csv', 'r') as source_file:
    with open('titanic3_new.csv', 'w') as dest_file:
        cnt = 0
        for line in source_file:
            new_line = line[1:-2] #get rid of quotes around the whole line
            new_line = new_line.replace('""', '') #get rid of all the other double quotes

                   if cnt == 0: #add id column
                new_line = "id," + new_line + '\n'
            else:
                new_line = str(cnt) + ',' + new_line + '\n'

            if cnt == 0: #split name column into name and surname columns
                new_line = new_line.replace('name,sex', 'name,surname,sex')
            else:
                new_line = re.sub(r'^(\d+,\d,\d,[^,]+), (.*)', r'\1,\2', new_line)

            #change all commas to semicolons -> directly viewable as a table in Excel
            new_line = new_line.replace(',', ';')
            #get rid of all semicolons within values by replacing them with a comma
            new_line = new_line.replace('; ', ', ')

            dest_file.write(new_line) #write new_line to destination file
            cnt += 1

print "done."
```

1.2 Data partitioning

The data set has further been divided into a training- and test set to assess the performance for your script; the partitioning has been conducted as follows (here for reference reasons):

```
import re

#change here to change the relative size of the test set compared to the training set
percent_test = 20

print "Partitioning data..."

with open('titanic3_new.csv', 'r') as source_file:
    with open('titanic3_train.csv', 'w') as train_file:
        with open('titanic3_test.csv', 'w') as test_file:
            cnt = 0
            mod_value = int(100.0 / percent_test)
            for line in source_file:
                if cnt == 0:
                    train_file.write(line)
                    new_line = line.replace('survived;', '') #omit 'survived' column
                    test_file.write(new_line)
                #decide if this line goes to the training- or test-set
                #(deterministically, not at random, so that it is repeatable)
                elif cnt % mod_value == 0:
                    #omit 'survived' column
                    new_line = re.sub(r'^(\d+;\d);(\d);(.*)', r'\1;\3', line)
                    test_file.write(new_line)
                else:
                    train_file.write(line)
                cnt += 1

print "Done."
```

Basically, after a header row containing the column names, each row contains data about one passenger. The training set **`titanic3_train.csv`** includes the information if he or she has survived; in the test data set **`titanic3_test.csv`**, this column has been omitted. See the following chapter for more background information and an explanation of the columns.

1.3 Original data description

NAME: titanic3
TYPE: Census
SIZE: 1309 Passengers, 14 Variables

DESCRIPTIVE ABSTRACT: The titanic3 data frame describes the survival status of individual passengers on the Titanic. The titanic3 data frame does not contain information for the crew, but it does contain actual and estimated ages for almost 80% of the passengers.

SOURCES: Hind, Philip. "Encyclopedia Titanica." Online. Internet.
n.p. 02 Aug 1999. Avaliable http://atschool.eduweb.co.uk/phind

VARIABLE DESCRIPTIONS:

id	a running number (starting from 1) to identify any record
pclass	Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
survived	Survival (0 = No; 1 = Yes)
name	Name
surname	Surname and title/salutation
sex	Sex
age	Age
sibsp	Number of Siblings/Spouses Aboard
parch	Number of Parents/Children Aboard
ticket	Ticket Number
fare	Passenger Fare
cabin	Cabin
embarked	Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)
boat	Lifeboat
body	Body Identification Number
home.dest	Home/Destination

SPECIAL NOTES:

Pclass is a proxy for socio-economic status (SES)
1st ~ Upper; 2nd ~ Middle; 3rd ~ Lower

Age is in Years; Fractional if Age less than One (1)
If the Age is Estimated, it is in the form xx.5

Fare is in Pre-1970 British Pounds (£)
Conversion Factors: 1£ = 12s = 240d and 1s = 20d

With respect to the family relation variables (i.e. sibsp and parch) some relations were ignored. The following are the definitions used for sibsp and parch.

Sibling: Brother, Sister, Stepbrother, or Stepsister of Passenger Aboard Titanic
Spouse: Husband or Wife of Passenger Aboard Titanic (Mistresses and Fiancées ignored)
Parent: Mother or Father of Passenger Aboard Titanic
Child: Son, Daughter, Stepson, or Stepdaughter of Passenger Aboard Titanic

Other family relatives excluded from this study include cousins, nephews/nieces, aunts/uncles, and in-laws. Some children travelled only with a nanny, therefore parch=0 for them. As well, some travelled with very close friends or neighbors in a village, however, the definitions do not support such relations.

STORY BEHIND THE DATA:
This dataset is based on the Titanic Passenger List edited by Michael A. Findlay, originally published in Eaton & Haas (1994) Titanic: Triumph and Tragedy, Patrick Stephens Ltd, and expanded with the help of the internet community. The original HTML files were obtained by Philip Hind (1999).

PEDAGOGICAL NOTES:
This dataset is ideal for teaching basic functions in S-PLUS in the realm of Statistical Computing and Graphics. It can also prove useful in teaching binary logistic regression and methods of imputation, both single and multiple. The dataset is also useful for demonstrating many of the functions available in Frank Harrell's Hmisc library as well as demonstrating binary logistic regression analysis using the Design library.

An interesting result may be obtained using functions from the Hmisc library in S-PLUS

```
attach(titanic3)
plsmo(age, survived, group=sex, datadensity=T)      # OR group=pclass
plot(naclus(titanic3))                      # study patterns of missing values
summary(survived ~ age + sex + pclass, data=titanic3)
```

REFERENCES:
Harrell FE. "Predicting Outcomes: Applied Survival Analysis and Logistic Regression." Book manuscript available from the University of Virginia Bookstore, 1999.

SUBMITTED BY:
Thomas E. Cason, Undergraduate Research Assistant
Division of Biostatistics and Epidemiology
Department of Health Evaluation Sciences
University of Virginia School of Medicine
Box 600, Charlottesville, VA 22908 USA
Electronic Mail: tcason@virginia.edu

FREQUENTLY ASKED QUESTIONS ABOUT THE DATASET

1. For those over age 25 the mean # spouses/siblings is about .34 - seems a little low

The only explanation I can offer (without a deep search) is the overwhelming "Third Class Bias" as I call it. Many third class passengers travelled alone... or some with friends... which is not under the umbrella of the sibsp definition. Also, many 3rd classers were immigrating to the US... they were married... but were sent off alone to establish a "foothole" and then later sent for their spouses... if they survived... most did not.

2. For those under age 14 the mean # parents/children is 1.37 - seems a bit low

Again... not all children travelled with their parents... especially in 3rd class. Some children travelled with older siblings... nannies... aunts/uncles... etc. Actually, more often
than not... children travelled with only one parent.

-TEC

After further investigation... I found my initial instincts regarding the low means to be correct. There's not much else to say about it... but I'll cite some unusual passenger cases that may come up in the future regarding this issue.

Case #1: Emanuel, Miss. Virginia Ethel... 3d Class... Age 5... sibsp/parch=0/0
Boarded with her nurse Miss. Elizabeth Dowdell... escorted her to grandparents' home in New York, NY.

Case #2: Hassan, Mr. Houssein G N... 3d Class... Age 11... s/p=0/0
Traveled with family friend Mr. Nassef Cassem Albimona... going to visit his parents in American from Lebanon. (Interesting Note: Albimona was from Fredericksburg, VA)

Case #3: Ayoub, Miss. Banoura... 3d Class... Age 13... s/p=0/0
Boarded with 5 cousins... travelling to Detroit, MI to be reunited with family.

Case #4: Nasser, Mrs. Nicholas Nasser... 2d Class... Age 14... s/p=1/0
Married to a 32 year old man... sibsp stands for spouse rather than sibling... unusual at such a young age. She lied when she boarded the Titanic and claimed she was 18... however, her birth certificate proves that on April 15, 1912 she was 14... not 18!

I hope this provides some insight to a few uncommon instances where the definitions do not encompass the actual travel status of a passenger.

There were only one or two instances of family members "crossing pclass lines"... and they were included and counted for in sibsp and parch.

-TEC