Lab: Hospital Data and Length of Stay (LOS)

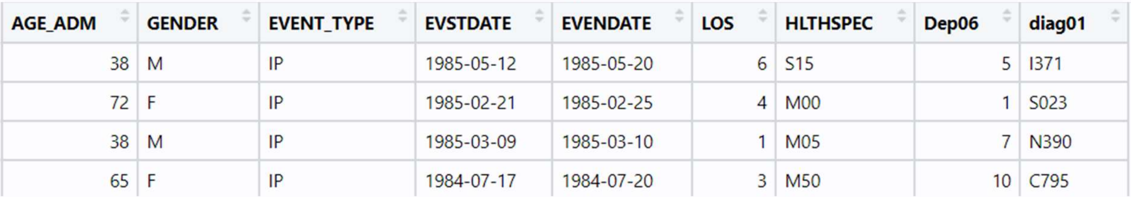
This lab introduces the hospital data which you will use for Assignment 4.

OBJECTIVES

By the end of this lab you will understand how to load in the data, explore some of the basic properties, build a simple model for LOS, and measure the quality of the prediction using the root relative mean squared error (RRSE).

DATA

cityA.csv, cityB.csv The two city datasets stored as CSV files.
Each dataset has the following set of columns:



AGE_ADM	GENDER	EVENT_TYPE	EVSTDATE	EVENDATE	LOS	HLTHSPEC	Dep06	diag01
38	M	IP	1985-05-12	1985-05-20	6	S15	5	I371
72	F	IP	1985-02-21	1985-02-25	4	M00	1	S023
38	M	IP	1985-03-09	1985-03-10	1	M05	7	N390
65	F	IP	1984-07-17	1984-07-20	3	M50	10	C795

See the lecture notes to explain the meaning of each of these columns. If you haven't seen the lecture material, please stop at this time and review the lecture.

CODE

bedcount.R This file has two functions –

bedcount <- function(data,LOScolumn=which(colnames(data)=="LOS"))

The *bedcount* function takes a dataset and returns a table with date and number of occupied beds.

plot.bc <- function(bc,r=1:nrow(bc),pts=TRUE,...)

The *plot.bc* function takes a table returned from "bedcount" and plots the data. The r parameter allows a subset of the timeseries to be displayed, and the pts variable (if TRUE) colours the points based on the day of the week.

GETTING STARTED

For purposes of demonstrating some concepts I'll just use city A.

a) Read in city A cA <- read.csv("cityA.csv")

How many examples are in this table (i.e. how many rows?). Ans: 66533

Exercise 1

Examine the LOS variable – do a histogram, min, max, etc. What are the characteristics of the length of stay for this city? What transformation may be useful for this variable? If you were going to model LOS are there some values you might exclude?

Is there any difference in LOS between the 2 genders (M and F)? Visualise and compare. NOTE: You may want to change some explanatory variables to factors.

Exercise 2:

Use the bedcount function to create a time series showing the bed counts over the year of data.

```
cA.bc <- bedcount(cA)
```

```
plot.bc(cA.bc)
```

```
plot.bc(cA.bc,r=150:200)
```

What are the patterns you observe with bedcount over time? Are there more or less people in hospital over the weekend? Examine the code in **bedcount.R** to understand how the weekday variable was created.

How would you create a new explanatory variable for the city data that was the weekday, and why might this be useful for modelling LOS?

The HLTHSPEC explanatory

What are the most common HLTHSPEC values? Look up a few of these in the supplied table (health_speciality_code.xls) and check what P60 and P70 means.

Exercise 3:

The HLTHSPEC for P60 and P70 relate to maternity services. Create a subset for cityA for just maternity HLTHSPEC's and check the LOS distribution. Is this what you might expect for someone going to hospital for the birth of a child?

Create the bed count data for this maternity subset and examine (plot) the behaviour. Would you expect that there is a pattern related to the weekend?

HYPOTHESIS: Giving birth is a random event. Show how might you answer (or at least provide some evidence) to confirm (or disprove) this hypothesis.

Exercise 4:

The diag01 explanatory variable has several thousand different values. What are the most common diagnoses for city A? Select one (have a look at the lecture notes for help) and create a subset of the data just for this initial diagnosis. A table to lookup these diagnoses may be found at:

https://icd.who.int/browse10/2019/en#

Once you have selected a diagnosis and made the subset of data, examine the LOS in comparison to the other explanatory variables. Do you think it will be possible to predict LOS for this diagnosis?

Exercise 5:

Open the file model.R and examine the functions. This file contains an example of function for calculating the root relative squared error (rrse), which is a good error measurement because values around 1 suggest the model is doing no better than predicting the mean of the response. A good model will have rrse values << 1.

The function

test.lm <- function(data,formula,perc.train=0.9)

{…}

takes a dataset and formula, and does a training/test split. The default is 90% training data, 10% test. Have a look at the code – it is just like the code we used for Assignment 2, however because we have factored variables these can be a problem with missing factors between the training and test data. The code does some extra work to stop the linear modelling from complaining. Of course there are other solutions to fix this problem (such as just catching the linear model when it fails and returning NA), but this is just a little more elegant….

Try predicting LOS for the single diagnosis data you built in ***Exercise 4***. Ensure that you have converted the appropriate explanatories to factors prior to running the model and any other transformations. Try this with different combinations of explanatories – which ones appear to help with prediction?

Finally – use replicate to test the model 50 times using what look like the "best" set of explanatory variables for predicting LOS. How well have you done? Try running a single model and get back the test (y) and predicted (yhat) values and examine the errors (HINT: Make a new function that is a modified version of test.lm, and return y and yhat as a table or a list). Are big LOS hard to predict, or are small values over/under predicted?