ex4

Nicole Navarro

September 19, 2016

The ability to predict discharge status using the available data will depend on which variables you plan to use. Different variables in the data have vastly differing levels of completeness. The completeness of variables can be examined using the following method.

We can check the number of cases which are complete across all variables, and we see that only about 1% of the cases are complete across all variables.

```
complete_cases<-complete.cases(diabetesData)
100*sum(complete_cases)/nrow(diabetesData)
```

```
## [1] 1.0249
```

We can also check for completeness in subsets of the data. For example we can check the completeness of admission type, discharge disposition, and admission source, and we see that 100% of the cases are complete with regards to these variables.

```
complete_cases<-complete.cases(diabetesData[,7:9])
100*sum(complete_cases)/nrow(diabetesData)
```

```
## [1] 100
```

So far we have only tested for complete cases with no missing data, but there are also situations where data was input that means the information is not available. For admission type, this is shown by responses containing NULL, Not Available, and Not Mapped. We can also check the percentage of the data not containing non-response data. We see that almost 90% of the data contains a response for this variable.

```
ad_type_clean<-which((diabetesData$admission_type_id!=5) & (diabetesData$admission_type_id !=6) & (diabe
clean_data<-diabetesData[ad_type_clean,]
100*nrow(clean_data)/nrow(diabetesData)
```

```
## [1] 89.78441
```

We can combine all of these techniques to examine the completeness of the different variables in the data and assess whether or not it is possible to create a prediction study. We can also use this information to see which variables should be included in the prediction study, and which are too irregular to be useful.