1 Building a simple data package for R

Suppose that we wish to make a package containing data sets only available in-house or on CRAN. This is often done for the data sets in the examples and exercises of a text book. For illustration, I will consider various data sets used in the Statistics Department's Masters exams over the years.

The simplest approach is to create the data files and then use the function package.skeleton to create the package. Even if you only make data sets available within your own research group, it will have the advantage that others can quickly install the package and have access to the data in the “usual” way.

2 Data files

2.1 bmd data

One of the simplest data files is bmd.dat in Fall 2003 master’s exam, stored as tab-separated values in the directory /p/stat/Data/MS.exam/f03/ on the AFS file system. We use the read.delim function to read the data. I usually check the structure immediately and decide if I need to post-process some of the columns.

```r
> dir <- "/p/stat/Data/MS.exam"
> str(bmd <- read.delim(file.path(dir, "f03", "bmd.dat")))
'data.frame': 126 obs. of 15 variables:
$ ID : Factor w/ 125 levels ":" ,"F02" ,"F03" ,...: 2 3 4 5 6 7 8 9 10 11 ...
$ Init : Factor w/ 123 levels "":"AES" ,"ALP" ,...: 93 84 81 95 94 64 45 65 30 92 ...
$ Age : int 64 78 59 74 51 68 57 57 83 50 ...
$ Gender : Factor w/ 3 levels ":" ,"F" ,"M": 2 2 2 2 2 2 2 2 2 2 ...
$ Height : num 161 169 169 163 168 ...
$ Weight : int 251 181 165 122 129 197 131 157 148 117 ...
$ L1.L4T : num 2.6 1.8 0.8 1.4 -1.6 -2.6 -1.6 1.8 -0.6 -2.3 ...
$ X : logi NA NA NA NA NA NA ...
$ INeckT : num 0.1 -1.0 0.1 0.2 -1.2 -0.5 0 1 -1.6 -1.3 ...
$ ITrochT: num 0.3 -0.6 0.6 -0.2 -2 -0.1 -1 1.9 -1.6 -0.8 ...
$ ITotalT: num 0.8 -0.8 0.2 0.1 -1.7 -0.5 -0.7 1 -1.8 -1.1 ...
$ DNeckT : num 0.1 -0.9 0.1 0.1 -1.4 -0.4 0.1 1.1 -1.6 -1.6 ...
$ DTrochT: num 0.6 -0.6 0.8 -0.1 -1.9 -0.2 -0.8 1.8 -1.5 -0.7 ...
$ DTotalT: num 0.8 -0.8 0.2 0.1 -1.7 -0.5 -0.7 1 -1.8 -1.1 ...
$ X.1 : logi NA NA NA NA NA NA ...

Immediately we can see some problems, although fewer problems than usually encountered. The factors all have a level of "" indicating that there are probably some completely blank lines in the data. Also there are two superfluous variables called "X" and "X.1" which are likely the result of extra tab characters on the lines.

First we handle the blank columns.

```r
> bmd <- bmd[, names(bmd)[!names(bmd) %in% c("X","X.1")]]

1
To assure ourselves that the lines with missing ID are indeed blank we consider the subset of the data frame corresponding to these rows.

```r
> subset(bmd, Gender=="")
```

```
ID Init Age Gender Height Weight L1.L4T INeckT ITrochT ITotalT DNeckT DTrochT
73 NA NA NA NA NA NA NA NA NA
74 NA NA NA NA NA NA NA NA NA

DTotalT
73 NA
74 NA
```

They are indeed blank in all the variables so we remove them

```r
> str(bmd <- subset(bmd, Gender != ""))
```

```
'data.frame': 124 obs. of 13 variables:
$ ID : Factor w/ 125 levels "","F02","F03",..: 2 3 4 5 6 7 8 9 10 11 ...
$ Init : Factor w/ 123 levels "","AES","ALP",..: 93 84 81 95 94 64 45 65 30 92 ...
$ Age : int 64 78 59 74 51 68 57 57 83 50 ...
$ Gender : Factor w/ 3 levels "","F","M": 2 2 2 2 2 2 2 2 2 2 ...
$ Height : num 161 169 165 122 129 197 131 157 148 117 ...
$ Weight : int 251 181 165 252 197 131 157 148 117 117 ...
$ L1.L4T : num 2.6 1.8 0.8 1.4 -1.6 -2.6 -1.6 1.8 -0.6 -2.3 ...
$ INeckT : num 0.1 -1 0.1 0.2 -1.2 -0.5 0 1 -1.6 -1.3 ...
$ ITrochT: num 0.3 -0.6 0.6 -0.2 -2 -0.1 -1 1.9 -1.6 -0.8 ...
$ ITotalT: num 0.8 -0.8 0.2 0.1 -1.7 -0.5 -0.7 1 -1.8 -1.1 ...
$ DNeckT : num 0.1 -0.9 0.1 0.1 -1.4 -0.4 0.1 1.1 -1.6 -1.6 ...
$ DTrochT: num 0.6 -0.6 0.8 -0.1 -1.9 -0.2 -0.8 1.8 -1.5 -0.7 ...
$ DTotalT: num 0.8 -0.8 0.2 0.1 -1.7 -0.5 -0.7 1 -1.8 -1.1 ...

Notice that the factors still have the blank level, even though we have eliminated the observations at that level. To clean things up we apply factor to those variables.

```r
> str(bmd <- within(bmd, + {
+   ID <- factor(ID)
+   Init <- factor(Init)
+   Gender <- factor(Gender)
+ })))
```

```
'data.frame': 124 obs. of 13 variables:
$ ID : Factor w/ 124 levels "F02","F03","F04",..: 1 2 3 4 5 6 7 8 9 10 ...
$ Init : Factor w/ 122 levels "AES","ALP","ALS",..: 92 83 80 94 93 63 44 64 29 9..
It looks like one of those “mixed units” studies where the height is recorded in cm. and the weight is recorded in pounds.

2.2 feltys data

Another straightforward data set is feltys.txt from fall 2005. This is also a tab-separated values file.
The `ScrSSN` column is a case identifier (presumably a scrambled social security number) and the `YearHosp` appears to be a year without the century, meaning that 2000 sorts before 1999. Consider the unique values

```
> xtabs(~ YearHosp, feltys)
YearHosp
   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19
 30  30  19  27  27 105 123  85  86  83  67  64  59  50  59  40  45  50  35  31
```

We repair those variables

```
> str(feltys <- within(feltys, 
+   { 
+     ScrSSN <- factor(ScrSSN) 
+     YearHosp <- ordered(ifelse(YearHosp < 85, 
+                     2000, 1900) + YearHosp) 
+   }))
'data.frame': 1115 obs. of 4 variables:
$ ScrSSN : Factor w/ 686 levels "662337","2861467",..: 17 32 42 47 79 85 92 104 1...
$ Gender : Factor w/ 2 levels "FEMALE","MALE": 2 2 2 2 2 2 2 2 2 2 ...
$ Age    : int 56 76 54 63 52 61 62 84 69 79 ...
$ YearHosp: Ord.factor w/ 20 levels "1985"<"1986"<..: 20 20 20 20 20 20 20 20 20 20 ...

> summary(feltys)
```
<table>
<thead>
<tr>
<th>ScrSSN</th>
<th>Gender</th>
<th>Age</th>
<th>Year</th>
<th>Hosp</th>
</tr>
</thead>
<tbody>
<tr>
<td>143196520</td>
<td>FEMALE</td>
<td>34</td>
<td>1986</td>
<td>123</td>
</tr>
<tr>
<td>545443700</td>
<td>MALE</td>
<td>1081</td>
<td>1985</td>
<td>105</td>
</tr>
<tr>
<td>40239436</td>
<td>MALE</td>
<td>62</td>
<td>1988</td>
<td>86</td>
</tr>
<tr>
<td>812344720</td>
<td>MALE</td>
<td>66.63</td>
<td>1987</td>
<td>85</td>
</tr>
<tr>
<td>3996116</td>
<td>MALE</td>
<td>73</td>
<td>1989</td>
<td>83</td>
</tr>
<tr>
<td>201004127</td>
<td>MALE</td>
<td>91</td>
<td>1990</td>
<td>67</td>
</tr>
</tbody>
</table>

(Other) : 1066 (Other): 566

> subset(feltys, ScrSSN == 143196520)

<table>
<thead>
<tr>
<th>ScrSSN</th>
<th>Gender</th>
<th>Age</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>170 143196520</td>
<td>MALE</td>
<td>74</td>
<td>1998</td>
</tr>
<tr>
<td>208 143196520</td>
<td>MALE</td>
<td>74</td>
<td>1997</td>
</tr>
<tr>
<td>303 143196520</td>
<td>MALE</td>
<td>72</td>
<td>1995</td>
</tr>
<tr>
<td>346 143196520</td>
<td>MALE</td>
<td>71</td>
<td>1994</td>
</tr>
<tr>
<td>402 143196520</td>
<td>MALE</td>
<td>70</td>
<td>1993</td>
</tr>
<tr>
<td>452 143196520</td>
<td>MALE</td>
<td>69</td>
<td>1992</td>
</tr>
<tr>
<td>513 143196520</td>
<td>MALE</td>
<td>68</td>
<td>1991</td>
</tr>
<tr>
<td>579 143196520</td>
<td>MALE</td>
<td>67</td>
<td>1990</td>
</tr>
<tr>
<td>648 143196520</td>
<td>MALE</td>
<td>65</td>
<td>1989</td>
</tr>
<tr>
<td>730 143196520</td>
<td>MALE</td>
<td>65</td>
<td>1988</td>
</tr>
</tbody>
</table>

An associated data set is

> str(raprev <- read.delim(file.path(dir, "f05", "raprev.txt")))

'data.frame': 20 obs. of 12 variables:
$ Hosp_Persons: int 366948 359247 359617 358573 353249 359608 371273 406556...
$ Hosp_Women: int 17134 15971 15740 15355 14488 14103 13741 14268 15974 1...
$ RA_Persons: int 3073 3091 3006 3037 2944 3037 3085 3291 3652 3923 ...
$ RA_Women : int 230 193 192 208 171 173 180 198 211 216 ...  
$ Felty_Persons: int 27 27 19 30 31 35 50 45 40 ...  
$ Felty_Women: int 0 1 1 1 0 1 1 3 0 0 ...  
$ Pleuritis_Persons: int 118 115 119 112 88 122 132 121 127 121 ...  
$ ILD_Persons: int 6 4 6 7 7 7 3 7 7 7 ...  
$ Vasculitis_Persons: int 69 64 64 72 156 160 200 203 184 192 ...  
$ Vasculitis_Women: int 7 8 2 5 2 6 9 9 7 4 ...  
$ Carditis_P: int 16 18 23 22 22 32 25 33 26 28 ...  

3 Producing a package skeleton

Now we produce a package skeleton.
> package.skeleton("WiscMSexam", c("bmd", "feltys", "raprev"))
> tree <- system("tree WiscMSexam", intern=TRUE)
> str(tree)

  chr [1:14] "WiscMSexam" "|-- data" "|-- bmd.rda" "|-- feltys.rda" ...

> cat(paste(tree, collapse="\n"), "\n")

WiscMSexam
|-- data
  |-- bmd.rda
  |-- feltys.rda
  `-- raprev.rda
|-- DESCRIPTION
|-- man
  |-- bmd.Rd
  |-- feltys.Rd
  |-- raprev.Rd
  `-- WiscMSexam-package.Rd
 `-- Read-and-delete-me

2 directories, 9 files