**Manipulation of data-frame data with dutility functions**

*Klaus Holst & Thomas Scheike*

*March 3, 2020*

---

**Simple data manipulation for data-frames**

- Renaming variables, Deleting variables
- Looking at the data
- Making new variables for the analysis
- Making factors (groupings)
- Working with factors
- Making a factor from existing numeric variable and vice versa

Here are some key data-manipulation steps on a data-frame which is how we typically organize our data in R. After having read the data into R it will typically be a data-frame, if not we can force it to be a data-frame. The basic idea of the utility functions is to get a simple and easy to type way of making simple data-manipulation on a data-frame much like what is possible in SAS or STATA.

The functions, say, dcut, dfactor and so on are all functions that basically does what the base R cut, factor do, but are easier to use in the context of data-frames and have additional functionality.

```r
library(mets)
data(melanoma)

# Check if melanoma is a data-frame
is.data.frame(melanoma)
```

```
[1] TRUE
```

Here we work on the melanoma data that is already read into R and is a data-frame.

**dUtility functions**

The structure for all functions is

- **dfunction(dataframe,y~x|ifcond,…)**

  to use the function on y in a dataframe grouped by x if condition ifcond is valid. The basic functions are
  - Data processing
  - dsort
• dreshape
• dcut
• drm, drename, ddrop, dkeep, dsubset
• drelevel
• dlag
• dfactor, dnumeric

Data aggregation
• dby, dby2
• dscalar, deval, daggregate
• dmean, dsd, dsum, dquantile, dcor
• dtable, dcount

Data summaries
• dhead, dtail,
• dsummary,
• dprint, dlist, dlevels, du nive

A generic function daggregate, daggr, can be called with a function as the argument
• daggregate(dataframe,y~x|ifcond,fun=function, . . . )
  without the grouping variable (x)
• daggregate(dataframe,~y|ifcond,fun=function, . . . )

A useful feature is that y and x as well as the subset condition can be specified using regular-expressions or by wildcards (default). Here to illustrate this, we compute the means of certain variables.

First just overall

```r
dmean(melanoma,~thick+I(log(thick)))
```

```
thick I(log(thick))
291.985366 5.223341
```

now only when days>500

```r
dmean(melanoma,~thick+I(log(thick))|I(days>500))
```

```
thick I(log(thick))
271.582011 5.168691
```

and now after sex but only when days>500

```r
dmean(melanoma,thick+I(log(thick))~~sex|I(days>500))
```
Manipulation of data-frame data with dutility functions

```r
sex thick I(log(thick))
1 0 242.958 5.060086
2 1 320.242 5.353321
```

and finally after quartiles of days (via the dcut function)

```r
dmean(melanoma, thick+I(log(thick))~I(dcut(days)))
```

```r
I(dcut(days)) thick I(log(thick))
[10,1.52e+03] 482.173 5.799525
(1.52e+03,2e+03] 208.549 4.987652
(2e+03,3e+03] 223.294 4.974759
(3e+03,5.56e+03] 250.196 5.120129
```

or summary of all variables starting with "s" and that contains "a"

```r
dmean(melanoma,"s*"+"a*"~sex|I(days>500))
```

<table>
<thead>
<tr>
<th>sex</th>
<th>status</th>
<th>days</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.831933</td>
<td>2399.143</td>
</tr>
<tr>
<td>1</td>
<td>1.714286</td>
<td>2169.800</td>
</tr>
</tbody>
</table>

**Renaming, deleting, keeping, dropping variables**

```r
melanoma=drename(melanoma, tykkelse~thick)
names(melanoma)
```

```
[1] "no" "status" "days" "ulc" "tykkelse" "sex"
```

**Deleting variables**

```r
data(melanoma)
melanoma=drm(melanoma, ~thick+sex)
names(melanoma)
```

```
[1] "no" "status" "days" "ulc"
```

**or sas style**

```r
data(melanoma)
melanoma=ddrop(melanoma, ~thick+sex)
names(melanoma)
```

```
[1] "no" "status" "days" "ulc"
```

**alternatively we can also keep certain variables**

```r
data(melanoma)
melanoma=dkeep(melanoma, ~thick+sex+status+days)
names(melanoma)
```

```
[1] "thick" "sex" "status" "days"
```

**This can also be done with direct assignment**

```r
data(melanoma)
ddrop(melanoma) <- ~thick+sex
names(melanoma)
```

```
[1] "no" "status" "days" "ulc"
```
Looking at the data

```r
data(melanoma)
dstr(melanoma)
```

'data.frame': 205 obs. of 6 variables:
$ no : int 789 13 97 16 21 469 685 7 932 944 ...
$ status: int 3 3 2 3 2 3 3 3 3 3 ...
$ days : int 10 30 35 99 185 204 232 232 279 279 ...
$ ulc : int 1 0 0 0 1 1 1 1 1 1 ...
$ thick : int 676 65 134 290 1208 484 516 1288 322 741 ...
$ sex : int 1 1 1 0 1 1 1 1 1 0 ...

The data can in Rstudio be seen as a data-table but to list certain parts of the data in output window

```r
dlist(melanoma)

no status days ulc thick sex
1 789 3 10 1 676 1
2 13 3 30 0 65 1
3 97 2 35 0 134 1
4 16 3 99 0 290 0
5 21 1 3909 1 706 1
---
201 317 2 4705 1 706 1
202 798 2 4668 0 612 0
203 806 2 4688 0 48 0
204 606 2 4926 0 226 0
205 328 2 5565 0 290 0
```

```r
dlist(melanoma, ~.|sex==1)

no status days ulc thick
1 789 3 10 1 676
2 13 3 30 0 65
3 97 2 35 0 134
5 21 1 3909 1 706
6 469 1 204 1 484
---
191 445 2 3909 1 806
195 415 2 4119 0 65
197 175 2 4207 0 65
198 493 2 4310 0 210
201 317 2 4492 1 706
```

```r
dlist(melanoma, ~ulc+days+thick+sex|sex==1)

ulc days thick sex
1 1 10 676 1
2 0 30 65 1
3 0 35 134 1
5 1 185 1208 1
6 1 204 484 1
---
191 1 3909 806 1
195 0 4119 65 1
197 0 4207 65 1
198 0 4310 210 1
201 1 4492 706 1
```

Getting summaries
MANIPULATION OF DATA-FRAME DATA WITH DUTILITY FUNCTIONS

```r

# Manipulation of data

```dsummary(melanoma)`

```ssummary(melanoma)

<table>
<thead>
<tr>
<th>no</th>
<th>status</th>
<th>days</th>
<th>ulc</th>
<th>thick</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>2.0</td>
<td>1.00</td>
<td>10</td>
<td>0.000</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>222.0</td>
<td>1st Qu.:1.00</td>
<td>1st Qu.:1525</td>
<td>1st Qu.:97</td>
</tr>
<tr>
<td>Median</td>
<td>469.0</td>
<td>Median :2.00</td>
<td>Median :2005</td>
<td>Median :194</td>
</tr>
<tr>
<td>Mean</td>
<td>463.9</td>
<td>Mean :1.79</td>
<td>Mean :2153</td>
<td>Mean :0.439</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>731.0</td>
<td>3rd Qu.:2.00</td>
<td>3rd Qu.:3042</td>
<td>3rd Qu.:356</td>
</tr>
<tr>
<td>Max</td>
<td>992.0</td>
<td>Max :3.00</td>
<td>Max :5565</td>
<td>Max :1.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
</tr>
<tr>
<td>1st Qu.</td>
</tr>
<tr>
<td>Median</td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>3rd Qu.</td>
</tr>
<tr>
<td>Max</td>
</tr>
</tbody>
</table>

# or for specific variables

```dsummary(melanoma,-thick+status+sex)`

```ssummary(melanoma,-thick+status+sex)`

<table>
<thead>
<tr>
<th>thick</th>
<th>status</th>
<th>sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.</td>
<td>10.0</td>
<td>Min. :1.00</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>97.0</td>
<td>1st Qu.:1.00</td>
</tr>
<tr>
<td>Median</td>
<td>194</td>
<td>Median :2.00</td>
</tr>
<tr>
<td>Mean</td>
<td>292</td>
<td>Mean :1.79</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>356</td>
<td>3rd Qu.:2.00</td>
</tr>
<tr>
<td>Max</td>
<td>1742</td>
<td>Max :3.00</td>
</tr>
</tbody>
</table>

# Summaries in different groups (sex)

```dsummary(melanoma,thick+days+status~sex)`

```ssummary(melanoma,thick+days+status~sex)`

<table>
<thead>
<tr>
<th>sex: 0</th>
<th>thick</th>
<th>days</th>
<th>status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.</td>
<td>10.0</td>
<td>Min. :99</td>
<td>Min. :1.000</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>97.0</td>
<td>1st Qu.:1636</td>
<td>1st Qu.:2.000</td>
</tr>
<tr>
<td>Median</td>
<td>162.0</td>
<td>Median :2059</td>
<td>Median :2.000</td>
</tr>
<tr>
<td>Mean</td>
<td>248.6</td>
<td>Mean :2283</td>
<td>Mean :1.833</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>306.0</td>
<td>3rd Qu.:3131</td>
<td>3rd Qu.:2.000</td>
</tr>
<tr>
<td>Max</td>
<td>1742.0</td>
<td>Max :5565</td>
<td>Max :3.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>sex: 1</th>
<th>thick</th>
<th>days</th>
<th>status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.</td>
<td>16.0</td>
<td>Min. :10</td>
<td>Min. :1.000</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>105.0</td>
<td>1st Qu.:1052</td>
<td>1st Qu.:1.000</td>
</tr>
<tr>
<td>Median</td>
<td>258.0</td>
<td>Median :1860</td>
<td>Median :2.000</td>
</tr>
<tr>
<td>Mean</td>
<td>361.1</td>
<td>Mean :1946</td>
<td>Mean :1.722</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>484.0</td>
<td>3rd Qu.:2784</td>
<td>3rd Qu.:2.000</td>
</tr>
<tr>
<td>Max</td>
<td>1466.0</td>
<td>Max :4492</td>
<td>Max :3.000</td>
</tr>
</tbody>
</table>

# and only among those with thin-tumours or only females (sex==1)

```dsummary(melanoma,thick+days+status-sex|thick<97)`

```ssummary(melanoma,thick+days+status-sex|thick<97)`

<table>
<thead>
<tr>
<th>sex: 0</th>
<th>thick</th>
<th>days</th>
<th>status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.</td>
<td>10.00</td>
<td>Min. :356</td>
<td>Min. :1.000</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>32.00</td>
<td>1st Qu.:1762</td>
<td>1st Qu.:2.000</td>
</tr>
<tr>
<td>Median</td>
<td>64.00</td>
<td>Median :2277</td>
<td>Median :2.000</td>
</tr>
<tr>
<td>Mean</td>
<td>51.48</td>
<td>Mean :2425</td>
<td>Mean :2.034</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>65.00</td>
<td>3rd Qu.:3185</td>
<td>3rd Qu.:2.000</td>
</tr>
<tr>
<td>Max</td>
<td>81.00</td>
<td>Max :4688</td>
<td>Max :3.000</td>
</tr>
</tbody>
</table>
sex: 1

<table>
<thead>
<tr>
<th>thick</th>
<th>days</th>
<th>status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.  : 16.00</td>
<td>Min. : 30</td>
<td>Min. : 1.000</td>
</tr>
<tr>
<td>1st Qu.: 30.00</td>
<td>1st Qu.:1820</td>
<td>1st Qu.:2.000</td>
</tr>
<tr>
<td>Median : 65.00</td>
<td>Median :2886</td>
<td>Median :2.000</td>
</tr>
<tr>
<td>Mean : 55.75</td>
<td>Mean :2632</td>
<td>Mean :1.875</td>
</tr>
<tr>
<td>3rd Qu.: 81.00</td>
<td>3rd Qu.:3328</td>
<td>3rd Qu.:2.000</td>
</tr>
<tr>
<td>Max. : 81.00</td>
<td>Max. :4207</td>
<td>Max. :3.000</td>
</tr>
</tbody>
</table>

\[\text{dsummary(melanoma, thick+status} \sim 1 | \text{sex==1})\]

<table>
<thead>
<tr>
<th>thick</th>
<th>status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.  : 16.0</td>
<td>Min. : 1.000</td>
</tr>
<tr>
<td>1st Qu.: 105.0</td>
<td>1st Qu.:1.000</td>
</tr>
<tr>
<td>Median : 258.0</td>
<td>Median :2.000</td>
</tr>
<tr>
<td>Mean : 361.1</td>
<td>Mean :1.722</td>
</tr>
<tr>
<td>3rd Qu.: 484.0</td>
<td>3rd Qu.:2.000</td>
</tr>
<tr>
<td>Max. : 1466.0</td>
<td>Max. :3.000</td>
</tr>
</tbody>
</table>

or

\[\text{dsummary(melanoma, -thick+status|sex==1)}\]

<table>
<thead>
<tr>
<th>thick</th>
<th>status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.  : 16.0</td>
<td>Min. : 1.000</td>
</tr>
<tr>
<td>1st Qu.: 105.0</td>
<td>1st Qu.:1.000</td>
</tr>
<tr>
<td>Median : 258.0</td>
<td>Median :2.000</td>
</tr>
<tr>
<td>Mean : 361.1</td>
<td>Mean :1.722</td>
</tr>
<tr>
<td>3rd Qu.: 484.0</td>
<td>3rd Qu.:2.000</td>
</tr>
<tr>
<td>Max. : 1466.0</td>
<td>Max. :3.000</td>
</tr>
</tbody>
</table>

To make more complex conditions need to use the I()

\[\text{dsummary(melanoma, thick+days+status} \sim \text{sex|I(thick<97 & sex==1))}\]

sex: 1

<table>
<thead>
<tr>
<th>thick</th>
<th>days</th>
<th>status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.  : 16.00</td>
<td>Min. : 30</td>
<td>Min. : 1.000</td>
</tr>
<tr>
<td>1st Qu.: 30.00</td>
<td>1st Qu.:1820</td>
<td>1st Qu.:2.000</td>
</tr>
<tr>
<td>Median : 65.00</td>
<td>Median :2886</td>
<td>Median :2.000</td>
</tr>
<tr>
<td>Mean : 55.75</td>
<td>Mean :2632</td>
<td>Mean :1.875</td>
</tr>
<tr>
<td>3rd Qu.: 81.00</td>
<td>3rd Qu.:3328</td>
<td>3rd Qu.:2.000</td>
</tr>
<tr>
<td>Max. : 81.00</td>
<td>Max. :4207</td>
<td>Max. :3.000</td>
</tr>
</tbody>
</table>

Tables between variables

\[\text{dttable(melanoma, -status+sex)}\]

<table>
<thead>
<tr>
<th>sex</th>
<th>0 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>28 29</td>
</tr>
<tr>
<td>2</td>
<td>91 43</td>
</tr>
<tr>
<td>3</td>
<td>7  7</td>
</tr>
</tbody>
</table>

All bivariate tables

\[\text{dttable(melanoma, -status+sex+ulc, level=2)}\]
manipulation of data frame data with dutility functions

status
sex 1 2 3
0 28 91 7
1 29 43 7

status
ulc 1 2 3
0 16 92 7
1 41 42 7

sex
ulc 0 1
0 79 36
1 47 43

All univariate tables

```r
dtable(melanoma,~status+sex+ulc,level=1)
```

status
sex 1 2 3
57 134 14

sex
0 1
126 79

ulc
0 1
115 90

and with new variables

```r
dtable(melanoma,~status+sex+ulc+dcut(days)+I(days>300),level=1)
```

status
sex 1 2 3
57 134 14

sex
0 1
126 79

ulc
0 1
115 90

```r
dcut(days)

[10,1.52e+03] (1.52e+03,2e+03] (2e+03,3.04e+03] (3.04e+03,5.56e+03]
52 51 51 51

I(days > 300)
FALSE TRUE
11 194
```

Sorting the data

To sort the data
```r
# Manipulation of data-frame data with dutility functions

data(melanoma)
mel <- dsort(melanoma, ~days)
dsort(melanoma) <- ~days
define variables for analysis
head(melanoma)

<table>
<thead>
<tr>
<th>no</th>
<th>status</th>
<th>days</th>
<th>ulc</th>
<th>thick</th>
<th>sex</th>
<th>thick2</th>
<th>lthick</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>789</td>
<td>3</td>
<td>10</td>
<td>1</td>
<td>676</td>
<td>1</td>
<td>456976</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>3</td>
<td>30</td>
<td>0</td>
<td>65</td>
<td>1</td>
<td>4225</td>
</tr>
<tr>
<td>3</td>
<td>97</td>
<td>2</td>
<td>35</td>
<td>0</td>
<td>134</td>
<td>1</td>
<td>4225</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>3</td>
<td>99</td>
<td>0</td>
<td>290</td>
<td>0</td>
<td>17956</td>
</tr>
<tr>
<td>5</td>
<td>21</td>
<td>1</td>
<td>185</td>
<td>1</td>
<td>1208</td>
<td>1</td>
<td>1459264</td>
</tr>
<tr>
<td>6</td>
<td>469</td>
<td>1</td>
<td>204</td>
<td>1</td>
<td>484</td>
<td>1</td>
<td>234256</td>
</tr>
</tbody>
</table>

and to sort after multiple variables increasing and decreasing

dsort(melanoma) <- ~days-status
define variables for analysis
head(melanoma)

<table>
<thead>
<tr>
<th>no</th>
<th>status</th>
<th>days</th>
<th>ulc</th>
<th>thick</th>
<th>sex</th>
<th>thick2</th>
<th>lthick</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>789</td>
<td>3</td>
<td>10</td>
<td>1</td>
<td>676</td>
<td>1</td>
<td>456976</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>3</td>
<td>30</td>
<td>0</td>
<td>65</td>
<td>1</td>
<td>4225</td>
</tr>
<tr>
<td>3</td>
<td>97</td>
<td>2</td>
<td>35</td>
<td>0</td>
<td>134</td>
<td>1</td>
<td>4225</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>3</td>
<td>99</td>
<td>0</td>
<td>290</td>
<td>0</td>
<td>17956</td>
</tr>
<tr>
<td>5</td>
<td>21</td>
<td>1</td>
<td>185</td>
<td>1</td>
<td>1208</td>
<td>1</td>
<td>1459264</td>
</tr>
<tr>
<td>6</td>
<td>469</td>
<td>1</td>
<td>204</td>
<td>1</td>
<td>484</td>
<td>1</td>
<td>234256</td>
</tr>
</tbody>
</table>

Making new variables for the analysis

To define a bunch of new covariates within a data-frame

data(melanoma)
melanoma = transform(melanoma, thick2=thick^2, lthick=log(thick))
define variables for analysis
head(melanoma)

<table>
<thead>
<tr>
<th>no</th>
<th>status</th>
<th>days</th>
<th>ulc</th>
<th>thick2</th>
<th>lthick</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>789</td>
<td>3</td>
<td>10</td>
<td>456976</td>
<td>6.516193</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>3</td>
<td>30</td>
<td>4225</td>
<td>4.174387</td>
</tr>
<tr>
<td>3</td>
<td>97</td>
<td>2</td>
<td>35</td>
<td>4225</td>
<td>4.174387</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>3</td>
<td>99</td>
<td>84100</td>
<td>5.669881</td>
</tr>
<tr>
<td>5</td>
<td>21</td>
<td>1</td>
<td>185</td>
<td>1459264</td>
<td>7.096721</td>
</tr>
<tr>
<td>6</td>
<td>469</td>
<td>1</td>
<td>204</td>
<td>234256</td>
<td>6.182085</td>
</tr>
</tbody>
</table>

When the above definitions are done using a condition this can
be achieved using the dtransform function that extends transform
with a possible condition

melanoma=dtransform(melanoma, ll=thick*1.05^ulc, sex==1)
melanoma=dtransform(melanoma, ll=thick, sex!=1)

define variables for analysis
head(melanoma)

<table>
<thead>
<tr>
<th>sex</th>
<th>ulc</th>
<th>ll</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>173.7342</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>197.3611</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>374.5532</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>523.1198</td>
</tr>
</tbody>
</table>
Making factors (groupings)

On the melanoma data the variable thick gives the thickness of the melanoma tumour. For some analyses we would like to make a factor depending on the thickness. This can be done in several different ways

```r
melanoma = dcut(melanoma, ~ thick, breaks=c(0, 200, 500, 800, 2000))
```

New variable is named thickcat.0 by default.
To see levels of factors in data-frame

```r
dlevels(melanoma)
```

```r
thickcat.0 # levels = 4
[1] "[0,200]" "(200,500]" "(500,800]" "(800,2e+03]"
```

Checking group sizes

```r
dtable(melanoma, ~ thickcat.0)
```

```
  thickcat.0
    [0,200] (200,500] (500,800] (800,2e+03]
    109   64    20     12
```

With adding to the data-frame directly

```r
dcut(melanoma, breaks=c(0, 200, 500, 800, 2000)) <- gr.thick1 ~ thick
dlevels(melanoma)
```

```r
thickcat.0 # levels = 4
[1] "[0,200]" "(200,500]" "(500,800]" "(800,2e+03]"
```

```r
gr.thick1 # levels = 4
[1] "[0,200]" "(200,500]" "(500,800]" "(800,2e+03]"
```

New variable is named thickcat.0 (after first cut-point), or to get quartiles with default names thick.cat.4

```r
dcut(melanoma) <- ~ thick # new variable is thick.cat.4
dlevels(melanoma)
```

```r
thickcat.0 # levels = 4
[1] "[10,97]
[97,194] "(194,356]" "(356,1.74e+03]"
```

```r
gr.thick1 # levels = 4
[1] "[10,97]" "(97,194]" "(194,356]" "(356,1.74e+03]"
```

or median groups, here starting again with the original data,

```r
data(melanoma)
dcut(melanoma, breaks=2) <- ~ thick # new variable is thick.2
dlevels(melanoma)
```
To control new names

```r
data(melanoma)
mela = dcut(melanoma, thickcat4+dayscat4~thick+days, breaks=4)
dlevels(mela)
```

This can also be typed out more specifically

```r
melanoma$gthick = cut(melanoma$thick, breaks=c(0, 200, 500, 800, 2000))
melanoma$gthick = cut(melanoma$thick, breaks=quantile(melanoma$thick), include.lowest=TRUE)
```

**Working with factors**

To see levels of covariates in data-frame

```r
data(melanoma)
dcut(melanoma, breaks=4) <- thickcat4~thick
dlevels(melanoma)
```

To relevel the factor

```r
dtable(melanoma, ~thickcat4)
melanoma = drelevel(melanoma, ~thickcat4, ref="(194, 356]")
dlevels(melanoma)
```
Manipulation of data-frame data with dutility functions

or to take the third level in the list of levels, same as above,

```r
melanoma = drelevel(melanoma, -thickcat4, ref=2)
dlevels(melanoma)
```

To combine levels of a factor (first combining first 3 groups into one)

```r
melanoma = drelevel(melanoma, -thickcat4, newlevels=1:3)
dlevels(melanoma)
```

or to combine groups 1 and 2 into one group and 3 and 4 into another

```r
dkeep(melanoma) <- ~thick+thickcat4
melanoma = drelevel(melanoma, gthick2~thickcat4, newlevels=list(1:2, 3:4))
dlevels(melanoma)
```

Changing order of factor levels
Manipulation of data-frame data with dutility functions

1. `dfactor(melanoma, levels=c(3,1,2,4)) <- thickcat4.2-thickcat4`
2. `dtable(melanoma, "thickcat4")`
3. `dtable(melanoma, thickcat4+thickcat4.2)`

```
thickcat4 #levels=:4
[1] "[10,97]" "(97,194]" "(194,356]" "(356,1.74e+03]"
-----------------------------------------

thickcat4.2 #levels=:4
[1] "(194,356]" "[10,97]" "(97,194]" "(356,1.74e+03]"
-----------------------------------------

Combining levels but now control factor-level names

1. `melanoma=drelevel(melanoma, gthick3~thickcat4, newlevels=list(group1.2=1:2, group3.4=3:4))`
2. `dlevels(melanoma)`

```
thickcat4.2 (194,356] [10,97] (97,194] (356,1.74e+03]

Making a factor from existing numeric variable and vice versa

A numeric variable "status" with values 1,2,3 into a factor by

1. `data(melanoma)`
2. `melanoma = dfactor(melanoma, ~status, labels=c("malignant-melanoma", "censoring", "dead-other"))`
3. `melanoma = dfactor(melanoma, sexl~sex, labels=c("females", "males"))`
4. `dtable(melanoma, ~sexl+status.f)`

```
status.f malignant-melanoma censoring dead-other
sexl
females 28 91 7
males 29 43 7

A gender factor with values "M", "F" can be converted into numerics by

1. `melanoma = dnumeric(melanoma, ~sexl)`
2. `dstr(melanoma, "sex*")`
3. `dtable(melanoma, ~'sex*', level=2)`
'data.frame': 205 obs. of 3 variables:
$ sex : int 1 1 1 0 1 1 1 1 0 0 ...
$ sexl : Factor w/ 2 levels "females","males": 2 2 2 1 2 2 2 2 1 1 ...
$ sexl.n: num 2 2 2 1 2 2 2 2 1 1 ...

<table>
<thead>
<tr>
<th>sex</th>
<th>females</th>
<th>males</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>126</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>79</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>sexl</th>
<th>females</th>
<th>males</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>126</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>79</td>
</tr>
</tbody>
</table>