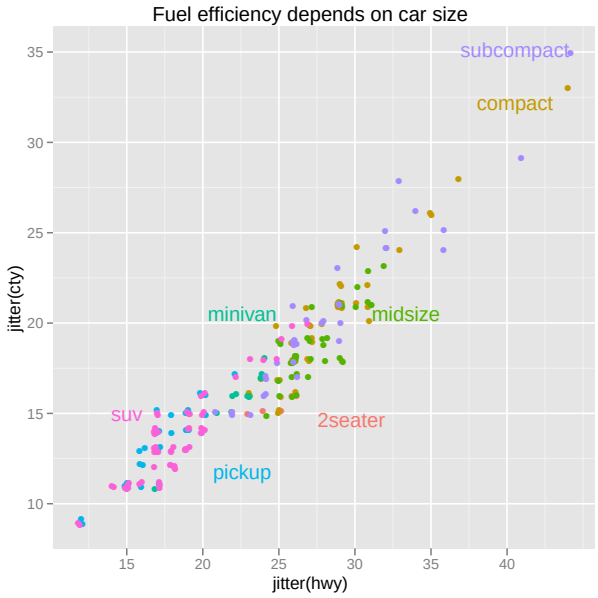
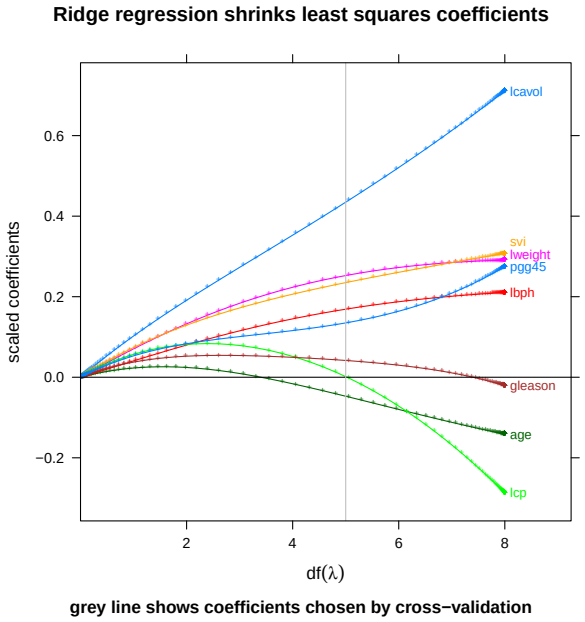
The directlabels package: easily add direct labels to R plots

Toby Dylan Hocking
toby.hocking@inria.fr
`http://cbio.ensmp.fr/~thocking`

January 26, 2010

Recent years have seen the development of sophisticated plotting systems for multivariate data in R. In particular, the power of the lattice and ggplot2 systems allows for automatic legend generation for plots labeled using different colors according to a categorical variable. This approach of using legends to decode color schemes is generally easy to implement but often makes for statistical graphics that are difficult or impossible to decode if there are too many colors. An easier method of decoding color schemes is by use of direct labels:





Direct labels are inherently more intuitive to decode than legends, since they are placed near the related data. However, direct labels are not widely used because they are often much more difficult to implement than legends, and their implementation varies between plotting systems.

The directlabels package solves these problems by providing a simple, unified interface for direct labeling in R. Given a lattice or ggplot2 plot saved in the variable `p`, direct labels can be added by calling `direct.label(p,f)` where `f` is a Positioning Function that describes where labels should be placed as a function of the data. The power of this system lies in the fact that you can write your own Positioning Functions, and that any Positioning Function can be used with any plot. So once you have a library of Positioning Functions, direct labeling becomes trivial and so can more easily used as a visualization technique in everyday statistical practice. The directlabels package comes with several Positioning Functions for different plot types, and a system of intelligent defaults so that often you don't even need to specify the Positioning Function.

The package was implemented using S3 methods to provide a unified interface for labeling lattice and ggplot2 plots, and could be extended to work with other plotting systems in R. This results in an elegant separation of the tasks of label position calculation and label drawing, thus resulting in a system that allows maximum customizability, code re-use, and user-friendlyness.

Development was started by Toby Dylan Hocking on 17 July 2009 while he was working toward his master's degree by doing a research internship at the French National Agriculture Institute (INRA) at Jouy-en-Josas. The research project he was working on called for the use of direct labels in several different contexts, but no general framework for doing this in R existed at the time. Since then, he has started doctoral studies in statistical learning with Jean-Philippe Vert and Francis Bach, which will be completed in 2012.

The package is currently hosted on R-Forge and can be installed using the following R code:

```
install.packages(c("ggplot2","ElemStatLearn","mlmRev"))#dependencies+examples
install.packages("directlabels",repos="http://r-forge.r-project.org")
```

There are 2 principal sources of documentation for the directlabels package. First, extensive documentation for all functions is included using the standard Rd mechanism for R packages. In particular, the plots shown above can be recreated by executing the R command `example(direct.label)` after loading the package with `library(directlabels)`. Second, a website hosted by R-Forge shows several examples that motivate the use of the package:

`http://directlabels.r-forge.r-project.org/`

Thank you for your consideration of the directlabels package for the John M. Chambers Statistical Software Award. The package's ability to make direct labeling easier will hopefully make clearer, direct labeled graphics more common in everyday statistical practice.