PLSC 502: "Statistical Methods for Political Research"

Bivariate and Multivariate Graphics
September 8, 2016

Overview

Examining more than one variable at a time graphically is the backbone of all good data analysis: from initial data exploration, through model specification and estimation to diagnostics, such plots should be and are used constantly. So, it's important to know how to do them well...

The key to understanding multivariate graphs is to think of them in terms of *conditionality*: When we plot more than one variable at a time, it is because we're interested in the *conditional* distribution of one variable (call it Y) as against another (call it X). In other words, rather than asking "How are literacy rates in Africa distributed?" we might as "How are literacy rates in *sub-saharan* Africa distributed?" or "How are literacy rates in Africa distributed *conditional on the level of wealth* in each country?"

As with univariate plots, what I'll discuss today are by no means the only options for plotting multivariate data. But they are the most widely-used methods, and so they're important to know.

Graph Types

We'll talk about three broad types of graphs today:

1. We'll spend a bunch of time on various forms of *scatterplots* for continuous data.
2. We'll then revisit *boxplots*, looking at them as a way of examining conditional distributions for discrete data.
3. Finally, we'll briefly discuss *contour plots* for three- (and higher-)dimensional data display.

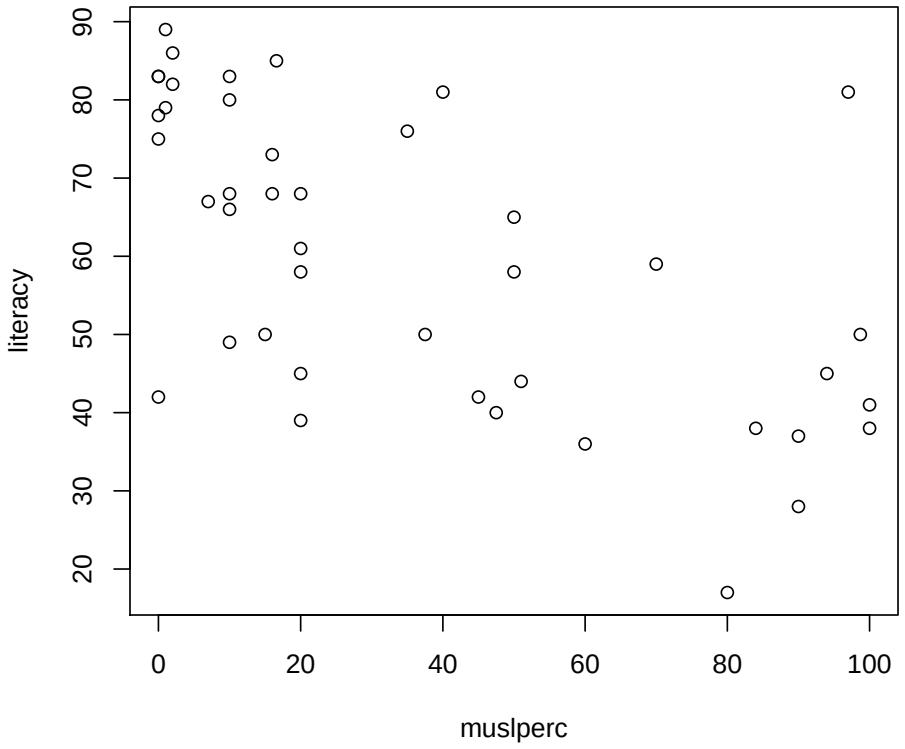
I'll organize the discussion around levels of measurement in one's data. That is, we'll talk about graphing various combinations of binary/nominal, ordinal, and interval/ratio variables against one another. Once again, plots will be generated using R . And once again, we'll use data on 43 African countries for the year 2001; they're (still) available on ANGEL.

Continuous Data

As the histogram is to univariate data display, so the *scatterplot* is to bivariate (and multivariate) data. A scatterplot just graphs two variables – one on the X-axis, one on the Y-axis – and plots each point in the data at their respective coordinate in that two–dimensional space:

```
> with(Africa, plot(muslperc,literacy))
```

Figure 1: Scatterplot of Literacy Rates and Muslim Population Percentages in Africa, 2001



This sort of plot can tell us quite a bit about the two variables in question. Here, we observe a generally downward trend: countries with higher Muslim populations have, on average, lower literacy rates than those with lower ones.

A few scatterplotting tips:

- As we'll see in a bit, scatterplots work best for relatively continuous data.
- Which variable is X and which is Y is up to you, thought the convention is that the "dependent" variable is measured along the Y-axis and the "independent" variable on the X axis.

- It's important to *label* the axes well; Stata generally does an OK job of this, but at times you'll need to supplement it.

There are also a lot of things one can do to make scatterplots more useful. These include:

- Using symbols that actually mean something (in instances where they do...).
- Including lines – either parallel to the X- or Y-axis – to indicate particularly important/salient values on that dimension.
- Adding "regression-type" lines and other indicators of the summary relationship between the two variables.

For our data here, then, we might replot the data like so:

```
> with(Africa, plot(muslperc,literacy,pch=19,ylab="Adult Literacy Rate",
                  xlab="Muslim Percentage of the Population"))
> with(Africa, text(muslperc,literacy,labels=cabbr,pos=3,cex=0.8))
> abline(h=mean(Africa$literacy,na.rm=TRUE),lty=2)
> abline(v=mean(Africa$muslperc,na.rm=TRUE),lty=2)
```

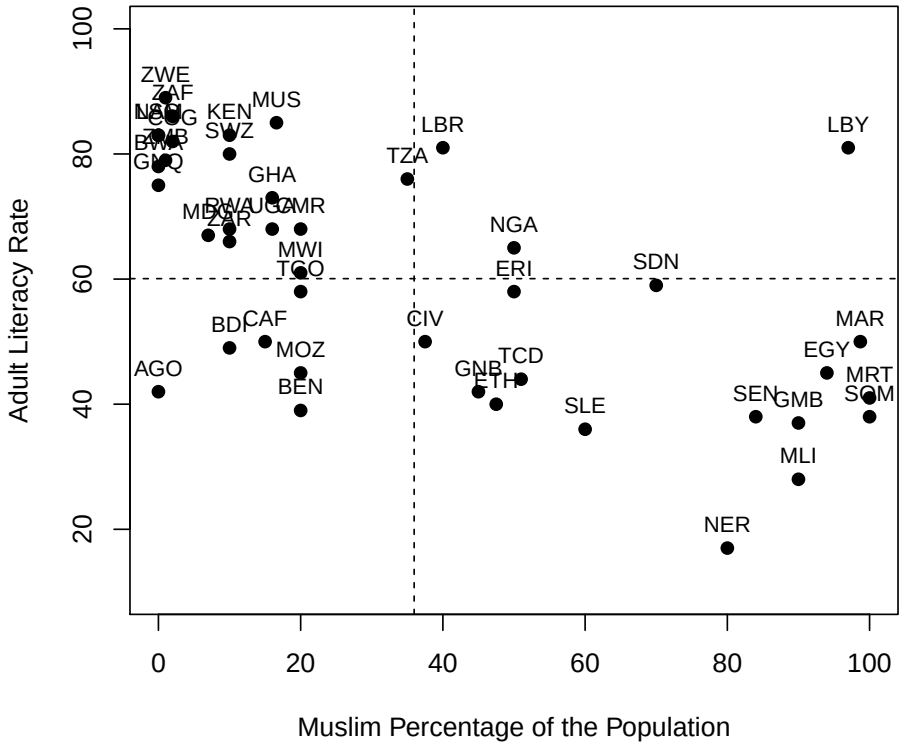
Note that this scatterplot includes indicators of the countries' names, as well as lines at the means of the two variables. It's just as useful as the one above, but also includes additional information.

Skewed Data

We often have variables that are skewed, particularly right-skewed (that is, with relatively few high-value observations); many economic variables, such as GDP, trade, etc. exhibit this characteristic. See what happens when we plot such variables:

```
> with(Africa, plot(gdppppd,tradegdp,pch=19,
     xlab="GDP Per Capita",ylab="Trade (% GDP)"))
```

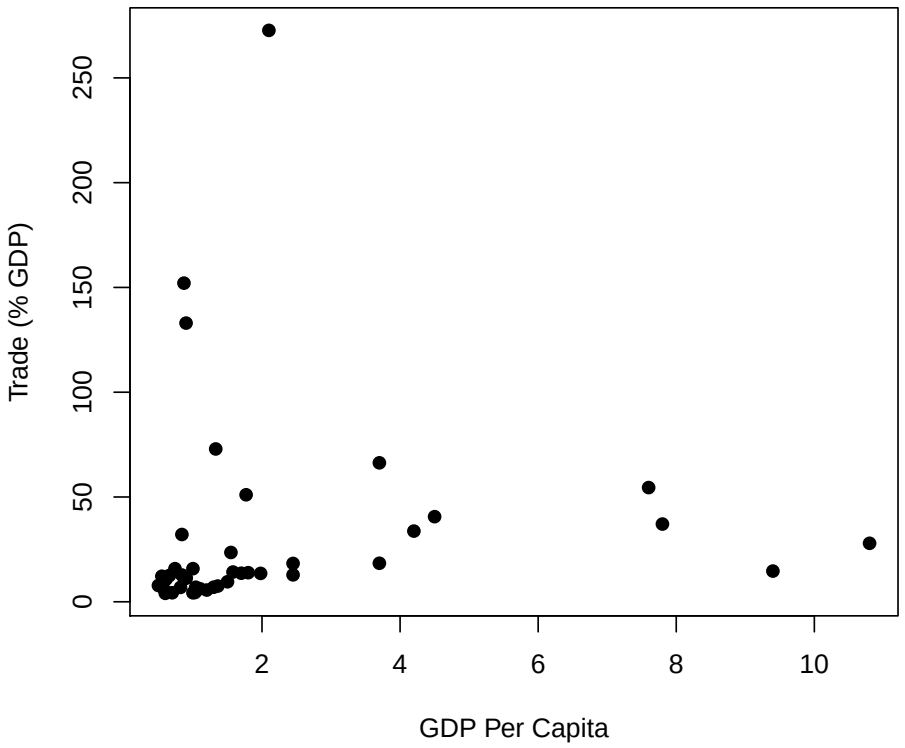
Figure 2: Revised Scatterplot of Literacy Rates and Muslim Population Percentages in Africa, 2001



This is not very useful at all; the scatterplot is dominated by a few "outliers."

A commonly-used way of dealing with this is to use a log-scale on the graph: instead of scaling the X and Y axes with the original values, we can instead plot (say) $\ln(Y)$ against $\ln(X)$:

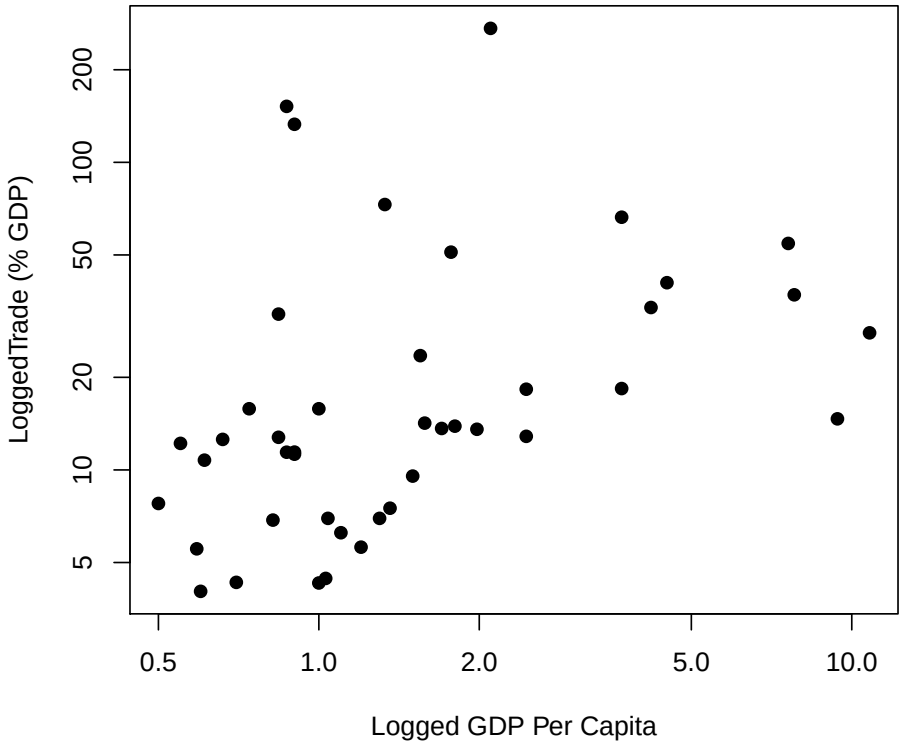
Figure 3: Scatterplot of Trade and GDP Per Capita in Africa, 2001




```
> with(Africa, plot(gdppppd,tradegdp,pch=19,log="xy",
                    xlab="Logged GDP Per Capita",
                    ylab="LoggedTrade (% GDP)"))
```

Note that now the scales are "stretched" at lower values of the variables, and "compressed" at higher values; the actual value labels remain correct. There's nothing strange or dishonest about this, provided that you make clear that you are using a log scale on the graph. We'll return to the issue of such data transformations later, when we talk about model specification and nonlinearity; for now, just remember that there are times when log scales can be valuable, particularly when data are positively (right-)skewed.

Figure 4: Log-Scale Scatterplot of Trade and GDP Per Capita in Africa, 2001



"Binned" Data

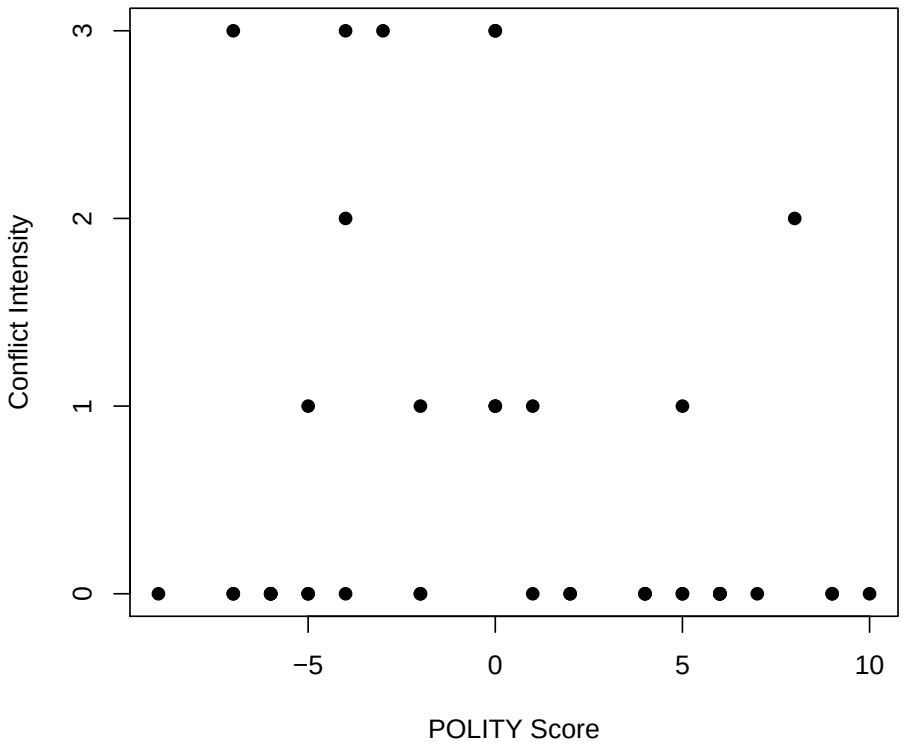
There are times when data are "binned;" that is, when variables don't have fully unique values. Sometimes this is a more serious problem than others; if there are relatively few data points with "tied" values, you can probably ignore it. However, in some instances, it can be a real problem for data display. Consider the variable `intensity`, which indicates the intensity of internal (civil) war in a country, as coded by the Uppsala/PRIO Armed Conflict Project:

- zero equals no conflict,
- one means low-level conflict (less than 25 deaths per year),
- two means intermediate conflict (more than 25 deaths per year, but less than 1000 deaths total), and
- three means high-level conflict (more than 1000 deaths per year).

Plotting this variable against the 21-point POLITY IV democracy indicator yields the following:

```
> with(Africa, plot(polity,intensity,pch=19,yaxp=c(0,3,3),
             xlab="POLITY Score",ylab="Conflict Intensity"))
```

Figure 5: Scatterplot of Domestic Conflict Intensity and POLITY Scores in Africa, 2001

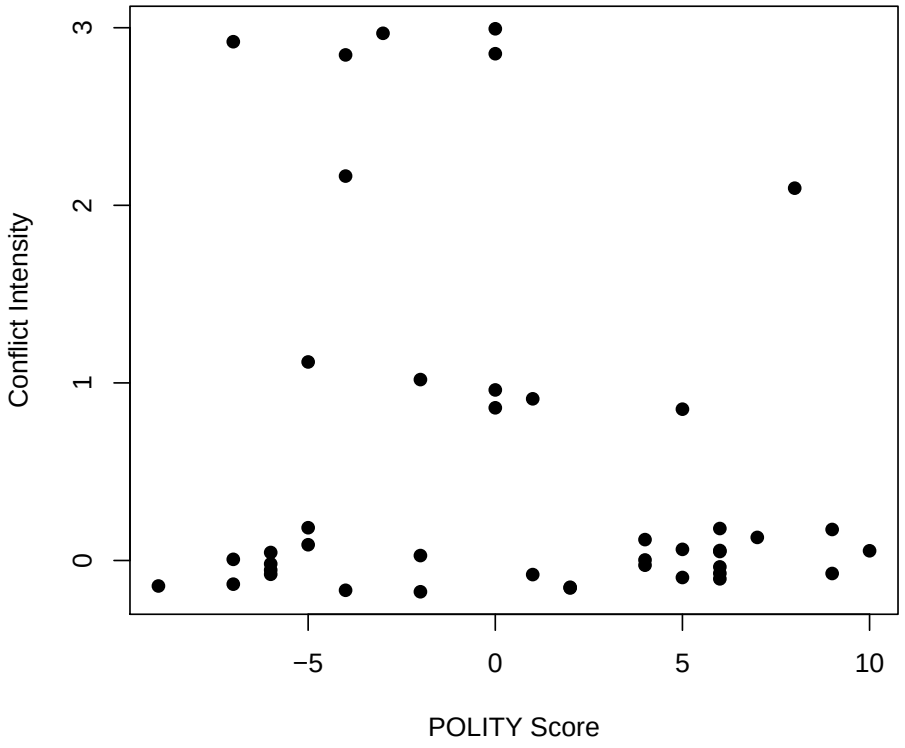


Note that several of the countries in the data have identical values on both of the variables; this can obscure the actual variation in the data (since it appears as if there is only one country where in fact there may be several).

One way to solve this issue is to introduce a bit of random variation ("jitter") into the scatterplot:

```
> set.seed(7222009)
> with(Africa, plot(polity,jitter(intensity,2),pch=19,yaxp=c(0,3,3),
                    xlab="POLITY Score",ylab="Conflict Intensity"))
```

Figure 6: Jittered Scatterplot of Domestic Conflict Intensity and POLITY Scores in Africa, 2001



This allows us to see that there are (e.g.) actually six countries, rather than five, that had low-intensity conflict in 2001. More generally, it gives us a better idea of "where" the data are located, without distorting the data.

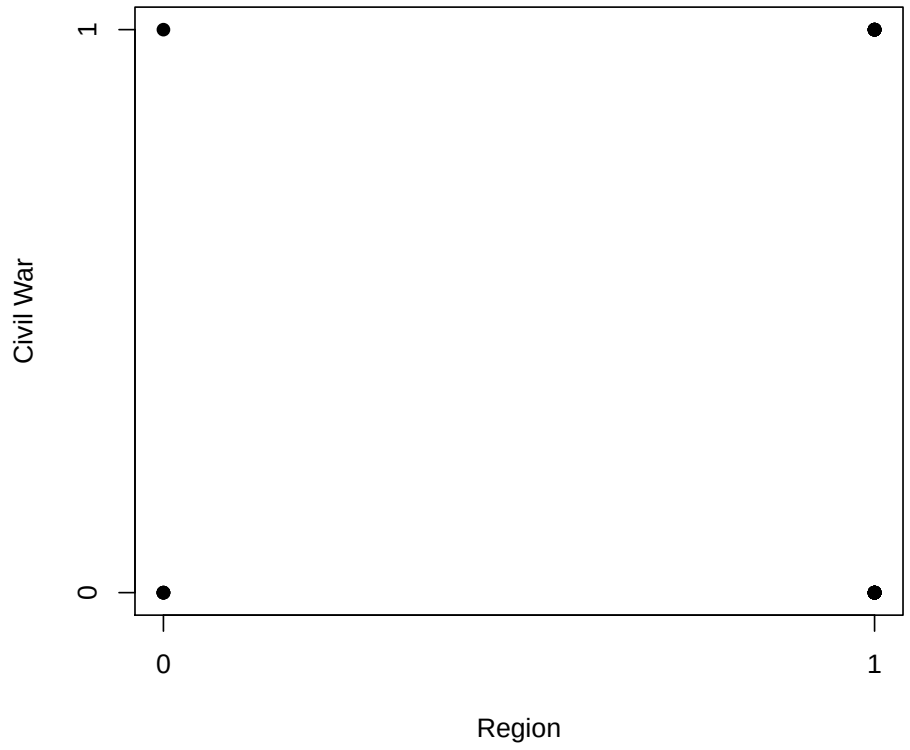
In fact, adding a bit of "jitter" to a scatterplot can be a good idea whenever the data are low-frequency ordinal. Moreover, this discussion of discrete data brings us to...

Binary Variables

Binary variables abound in political science, in part because you can take just about any other variable and make it binary (albeit with some loss of information). That said, binary variables present some real challenges for graphical display. For example, if you were interested in the relationship between (say) civil wars and their location in Africa, you might think to plot a measure of the presence of civil war (`internalwar`) against whether (=1) or not (=0) the country was in `subsaharan` Africa:

```
with(Africa, plot(as.numeric(subsaharan)-1,internalwar,pch=19,
                  xaxp=c(0,1,1),yaxp=c(0,1,1),xlab="Region",
                  ylab="Civil War"))
```

Figure 7: How Not To Draw a Scatterplot



In fact, what you really want to know is the frequency of `internalwar` for the two values of `subsharan`; in this case, we have one of the few instances where a table is actually better than a figure:

```
> with(Africa, xtabs(~subsaharan+internalwar))

                internalwar
subsaharan        0  1
  Not Sub-Saharan 5  1
  Sub-Saharan    25 12
```

Internal conflict appears to be somewhat more common in sub-Saharan Africa than above the Sahara. More to the point, a scatterplot of two binary variables is *never* a good idea.

Mixed Binary-Continuous Plots

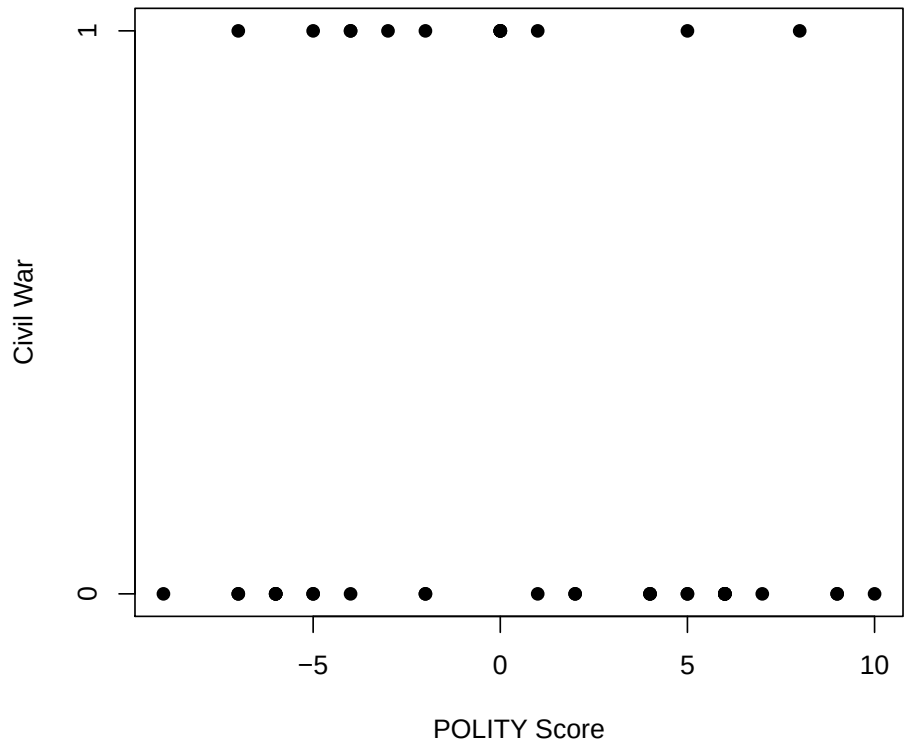
It is often the case that we want to examine variables of different types: particularly common are instances where we either want to know:

- How some binary variable Y covaries with some continuous variable X, or
- How some binary variable Y is different for different values of some binary variable X.

In the case of the binary dependent variable, it is possible to use a standard scatterplot, albeit with some modifications. A basic scatterplot leaves a lot to be desired:

```
with(Africa, plot(polity,internalwar,pch=19,
                  yaxp=c(0,1,1),xlab="POLITY Score",
                  ylab="Civil War"))
```

Figure 8: Scatterplot of Internal Conflict by POLITY Score for Africa, 2001

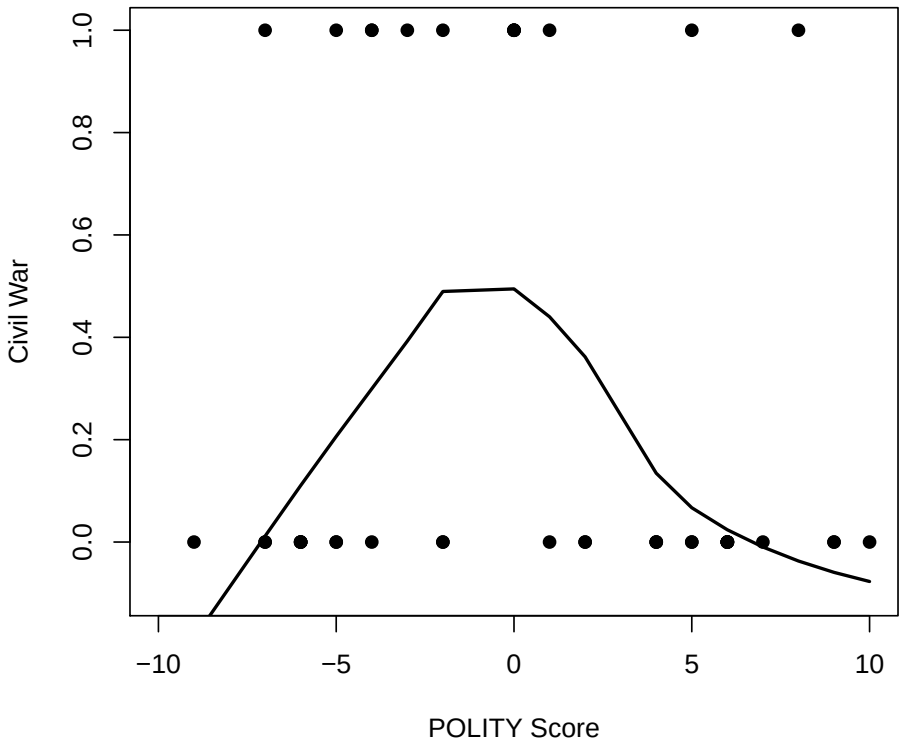


A better approach is to add a smoothed representation of the relationship, one that describes the "density" of the data at the various points on the X-axis. We can do this in several ways, but a good general approach is through adding a lowess (stands for "**lo**cally **we**ighted regre**ss**ion") line to the data. While there are a number of ways of doing lowess regression, the important intuition is that the lowess line represents something like the predicted value of the Y-axis variable at and around (that is, conditional on) that value of X.

Practically speaking, such a figure looks like this:

```
par(mar=c(4,4,2,2))
with(Africa, plot(lowess(polity,internalwar),xlab="POLITY Score",
                  ylab="Civil War",t="l",lwd=2,ylim=c(-0.1,1)))
with(Africa, points(polity,internalwar,pch=19))
legend(off)
```

Figure 9: Scatterplot and Lowess Regression of Internal Conflict by POLITY Score for Africa, 2001

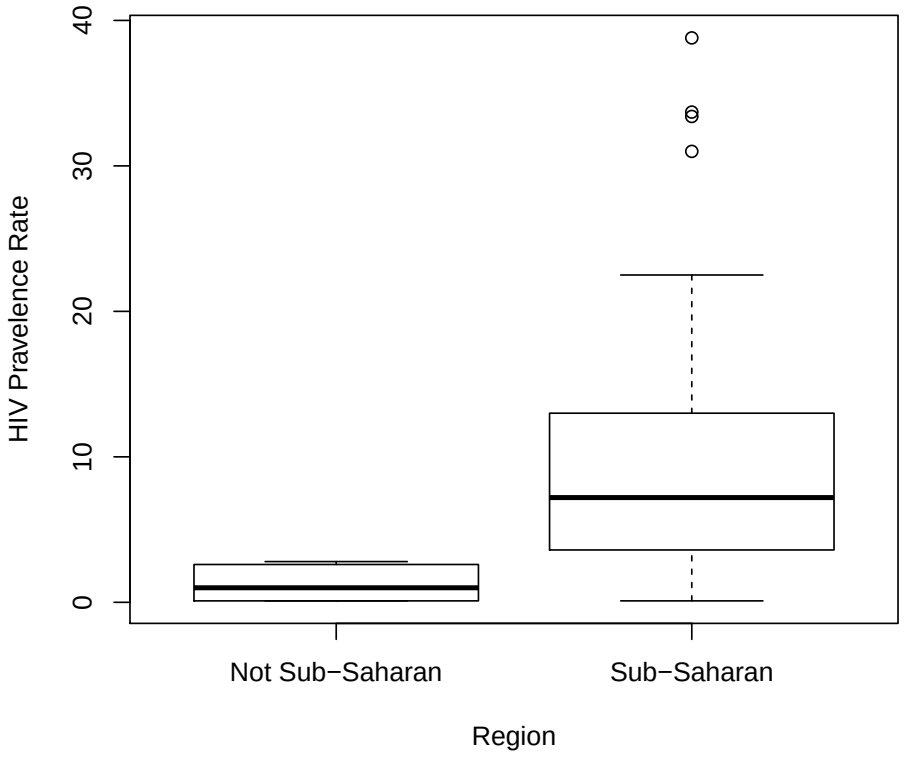


The line gives an idea of the general "shape" of the data: here, it tells us that civil wars apparently are more common at "medium" values of POLITY (that is, they are relatively rare among both strongly democratic and strongly autocratic systems, e.g. Gleditsch et al. 2001, de Soysa 2002, etc.).

When the binary variable is the "independent" variable, and the measure of the thing we are really interested in is continuous, we need to adopt another approach. Here, we can return to the boxplots we talked about last time; but this time, we can generate different boxplots of Y for different values of the (binary) independent variable X. So, if we wanted to see if there were any observable differences in adult HIV/AIDS prevalence rates between Saharan and sub-Saharan Africa, we could plot them:

```
with(Africa, boxplot(adrate~subsaharan,xlab="Region",
                     ylab="HIV Pravelence Rate"))
```

Figure 10: Boxplots of HIV/AIDS Rates by Region of Africa, 2001



This tells us that sub-Saharan Africa:

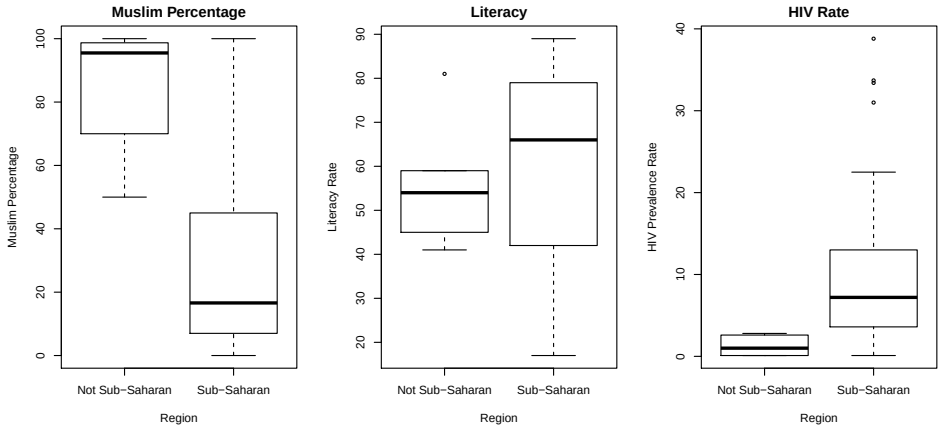
- Has, on average, higher HIV/AIDS rates than Saharan Africa, and
- also has greater variation in infection rates as well.[1]

More generally, we could plot several such variables, were we interested in comparing them all at once:

[1]None of that will come as any surprise to anyone who knows anything about AIDS in Africa.

```
> par(mfrow=c(1,3))
> with(Africa, boxplot(muslperc~subsaharan,xlab="Region",
            ylab="Muslim Percentage",cex=0.6,
            main="Muslim Percentage"))
> with(Africa, boxplot(literacy~subsaharan,xlab="Region",
            ylab="Literacy Rate",cex=0.6,
            main="Literacy"))
> with(Africa, boxplot(adrate~subsaharan,xlab="Region",
            ylab="HIV Prevalence Rate",cex=0.6,
            main="HIV Rate"))
```

Figure 11: Boxplots of Muslim Population, Literacy Rates, and HIV/AIDS Rates by Region of Africa, 2001



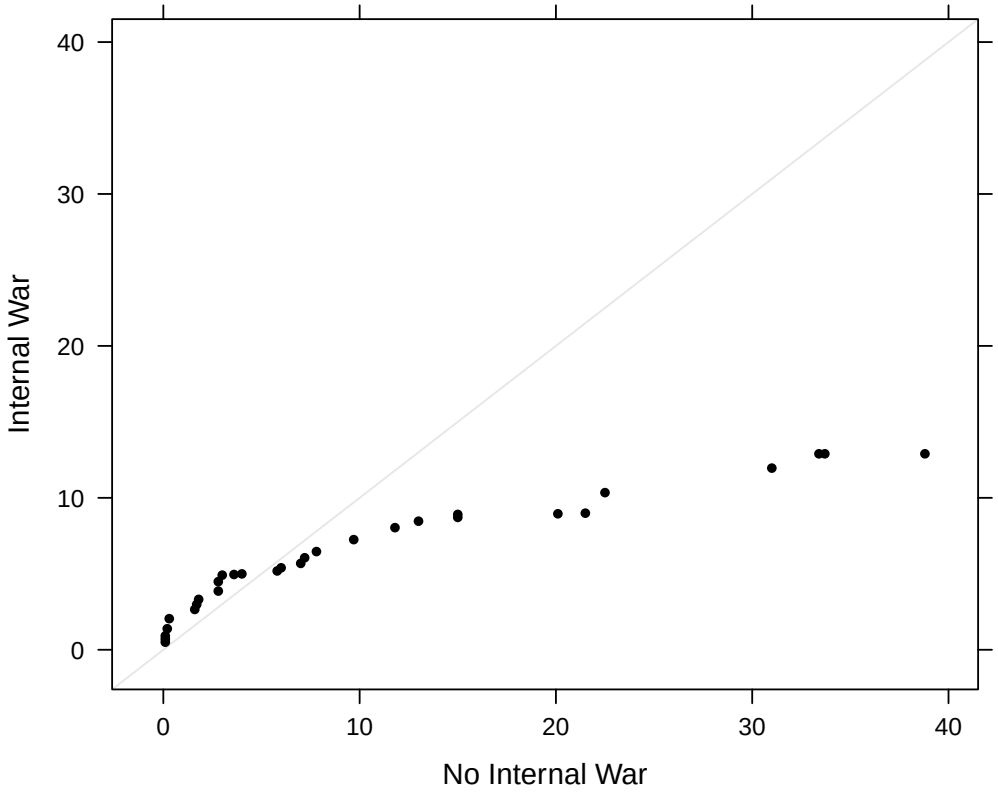
Q-Q Plots of Two Variables/Groups

We discussed Q-Q plots for comparing the empirical distribution of some data to a theoretical distribution. We can also use Q-Q plots for bivariate comparisons, either (a) to compare the distributions of two different variables in our data, or (b) to compare the distributions of some variable for two different groups in our data. The idea is that – rather than comparing our data of interest with the quantiles of some (known or hypothesized) theoretical distribution, we instead compare the quantiles for two different variables (or groups) in our data to each other.

For example, suppose we want to see if the distribution of HIV rates in countries with internal wars is similar to that for countries that did not have such a war. We do that by:

```
> library(lattice)
> with(Africa, qq(internalwar~adrate, col="black",pch=20,
            xlab="No Internal War", ylab="Internal War"))
```

Figure 12: Q-Q Plot of HIV Rates, Countries with and Without Internal Wars



This sorts the values of `adrate` within groups defined by`internalwar` and then plots them against each other. The plot indicates that the distribution of HIV rates is significantly different for the two groups; there is a much larger "spread" of values of `adrate` among countries for which `internalwar = 0` than among those where `internalwar = 1`. This is something we might have figured out any number of ways, actually, including via conditioned boxplots like the previous example.

Graphing Multivariate Data

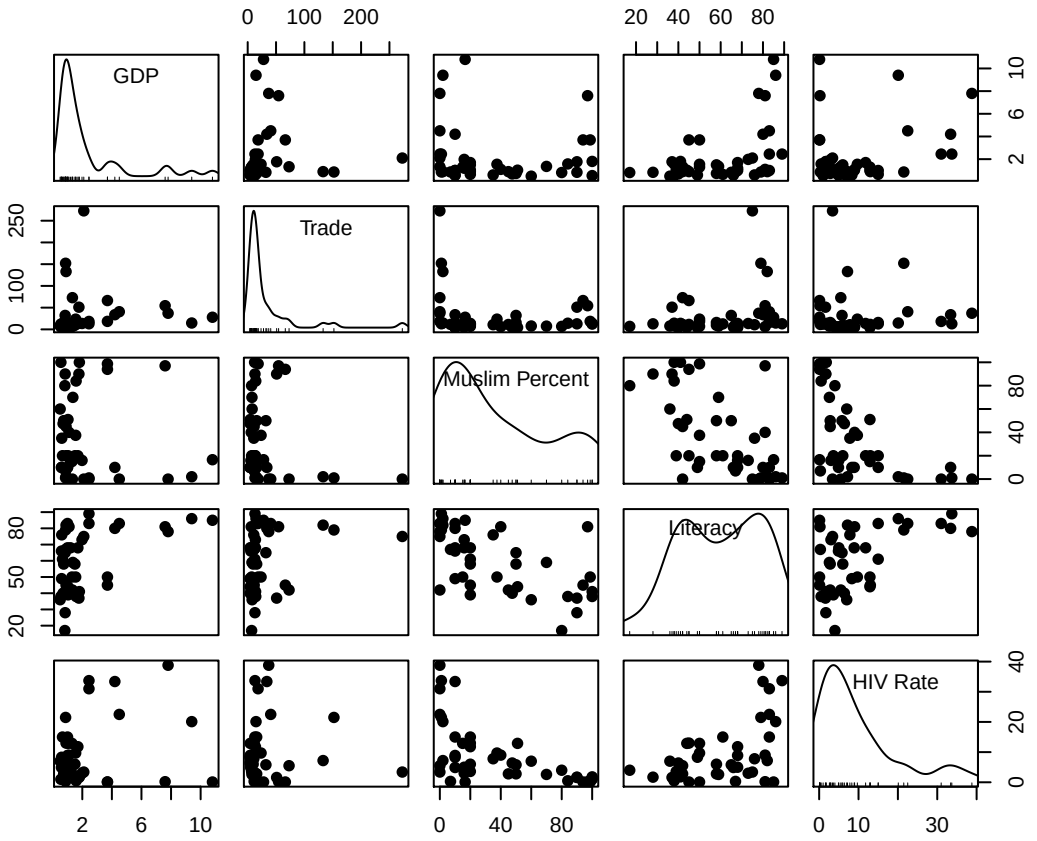
Here, I refer to data involving three or more variables as "multivariate" data. As Fox notes, it's not really natural for us to visualize more than two variables. Nonetheless, there are a few ways that we can do so. All are based on the idea of multiple conditioning, and all involve compromises.

Scatterplot Matrices

If we want to look at a number of variables plotted against each other all at once, we can use a *scatterplot matrix*. This is exactly what it sounds like: a bunch of scatterplots, arranged into matrix-like form. They look like this:

```
> library("car")
> dd <- Africa[,c("gdppppd","tradegdp","muslperc",
                  "literacy","adrate")]
> scatterplotMatrix(dd,reg.line=FALSE,smoother=FALSE,pch=19,
     var.labels=c("GDP","Trade","Muslim Percent","Literacy","HIV Rate"))
```

Figure 13: Scatterplots of GDP, Trade, Muslim Population, Literacy Rates, and HIV/AIDS Rates in Africa, 2001



Note a couple things about this:

- The scales are all standard/linear; if you want log-scales, you have to create a logged variable and include it in the variable list.
- Other than that, though, the scales adjust automatically to the scales of the variables (which is nice).

Scatterplot matrices are a really nice way to get an overall picture of your data the first time you check it out. What they *can't* show you is how more than two variables covary simultaneously...

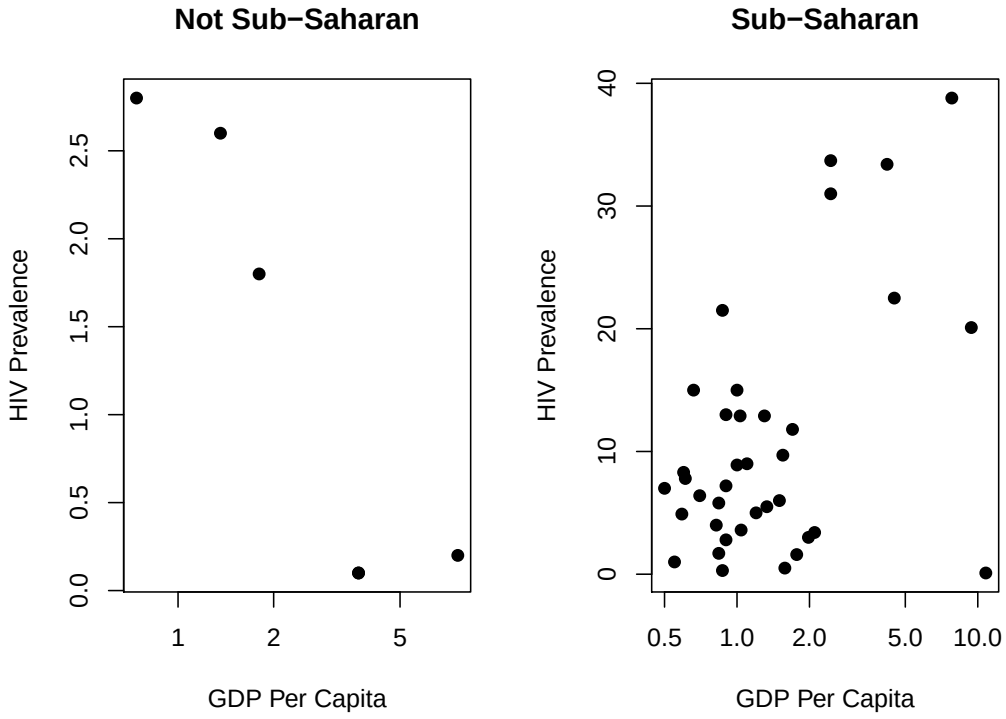
Conditional Plots

As we did with the boxplots above, we can examine more than two variables by conditioning bivariate plots on one or more other variables; this works best if at least one of the ("independent") variables is dichotomous, or at least discrete. For example, we might have reason to believe that the association between development (measured as GDP per capita) and HIV/AIDS rates might be different for Saharan countries than for sub-Saharan nations. To see if this is the case, we can generate conditional plots – scatterplots of two variables where we only include a subset of the cases in the data in each:

```
> par(mfrow=c(1,2)) # <- Create a combined plot: 1 row, 2 columns
> with(Africa[Africa$subsaharan=="Not Sub-Saharan",],
     plot(gdppppd,adrate,pch=19,main="Not Sub-Saharan",log="x",
     xlab="GDP Per Capita",ylab="HIV Prevalence"))
> with(Africa[Africa$subsaharan=="Sub-Saharan",],
     plot(gdppppd,adrate,pch=19,main="Sub-Saharan",log="x",
          xlab="GDP Per Capita",ylab="HIV Prevalence"))
```

Here, it appears as if the relationship between wealth and HIV/AIDS rates is negative (as would be expected) in Saharan Africa, but positive in sub-Saharan Africa. That's interesting to know.

Figure 14: Scatterplot of GDP and HIV/AIDS Rates, By Region, 2001



Contour Plots, Surface Plots, and Other 3-D Adventures

Finally, suppose you are really interested in knowing how three variables covary: say, What is the relationship among literacy, the Muslim population percentage, and HIV/AIDS rates? We can see from the scatterplot matrix above that

- countries with higher Muslim populations tend to have lower HIV/AIDS rates,
- countries with higher literacy rates tend to have higher HIV/AIDS rates, and
- countries with higher Muslim populations tend to have lower literacy rates.

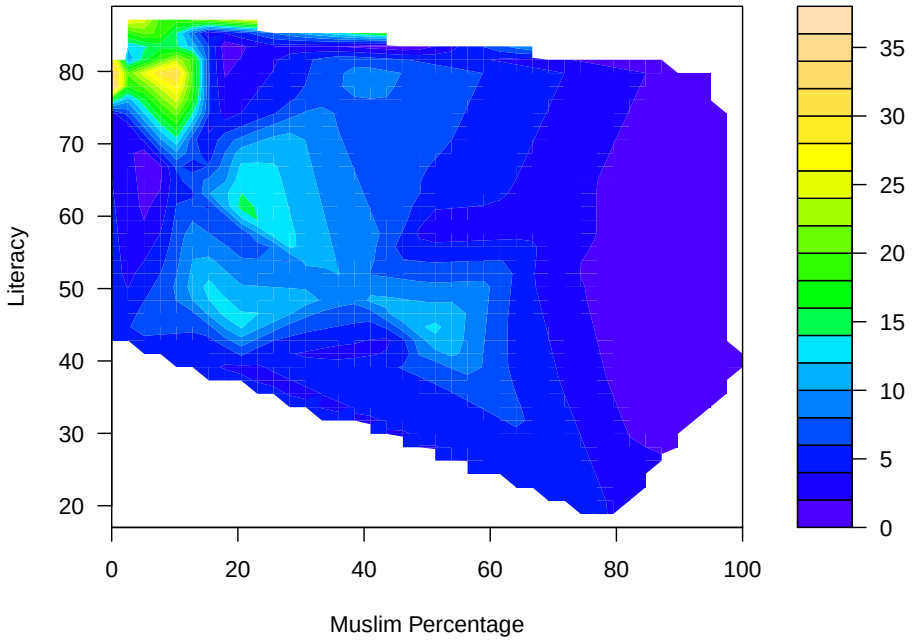
If we want to get a picture of all three variables at once, we have a couple options – neither of which, unfortunately, are currently implemented in Stata, or at least not very well.

Probably the best option is a *contour plot* – a representation of a three-dimensional graph in two dimensions. Think about a topographical map: it represents two variables/directions (latitude and longitude) as well as a third (elevation); the former are indicated by their locations in the X-Y space, while the latter are labeled:

```
library(akima)
```

```
cpdata <- with(Africa, interp(muslperc,literacy,adrate,
                        duplicate="mean"))
filled.contour(cpdata,color.palette=topo.colors,
                xlab="Muslim Percentage",
                ylab="Literacy")
```

Figure 15: Contour Plot of Muslim Population, Literacy Rates, and HIV Rates in Africa, 2001

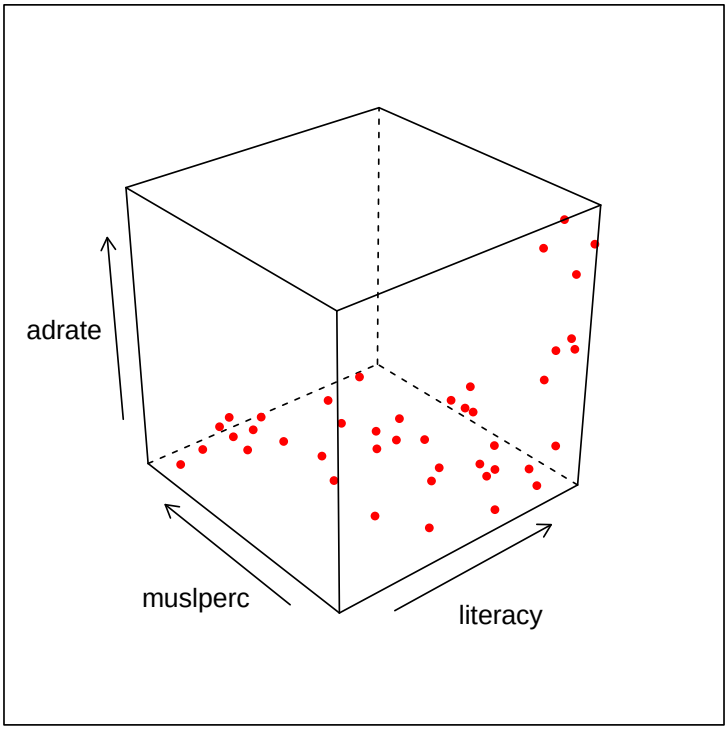


Note:

- The contours tell us the level of HIV/AIDS in those "regions" defined by the contour lines. This means that
- In general, we see the highest levels of HIV/AIDS rates in countries with high literacy rates and low Muslim populations (i.e., those in the upper-left corner of the figure).

Contour plots are (IMO) generally preferable to other sorts of plots for three-dimensional data. That said, another alternative is a *three-dimensional scatterplot*, which presents a two-dimensional representation of what would be a 3-D plot of the data:

Figure 16: Three-Dimensional Scatterplot of Muslim Population, Literacy Rates, and HIV/AIDS Rates in Africa, 2001



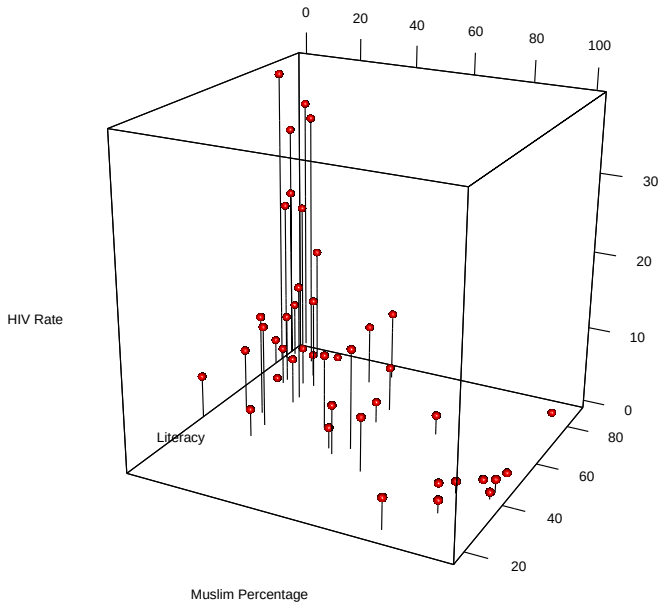
This scatterplot gives a really stark indication of the relationship among the three variables; the large "spikes" (high HIV/AIDS rates) occur at the "back-right corner" of the figure, at the place where Muslim population percentages are low and literacy rates are (relatively) high. This plot was done using `cloud`; one can also use the `scatterplot3d` function to do such plots.

For even more fun with these plots, the `rgl` package allows you to do real-time 3-D rendering of such plots, and creates interactive plots that one can rotate with the mouse:

```
> library(rgl)
> with(Africa, plot3d(muslperc,literacy,adrate,
              size=0.8, col="red",type="s",
              xlab="Muslim Percentage",
              ylab="Literacy",zlab="HIV Rate"))
> with(Africa, plot3d(muslperc,literacy,adrate,
     size=1, type="h",xlab="Muslim Percentage",
     ylab="Literacy",zlab="HIV Rate",add=TRUE)) # (Add lines)
```

```
> rgl.postscript("InteractiveMuslimLiteracyHIVScatter.pdf",
                 fmt="pdf") # (Save the output)
```

Figure 17: Interactive Three-Dimensional Scatterplot of Muslim Population, Literacy Rates, and HIV Rates in Africa, 2001

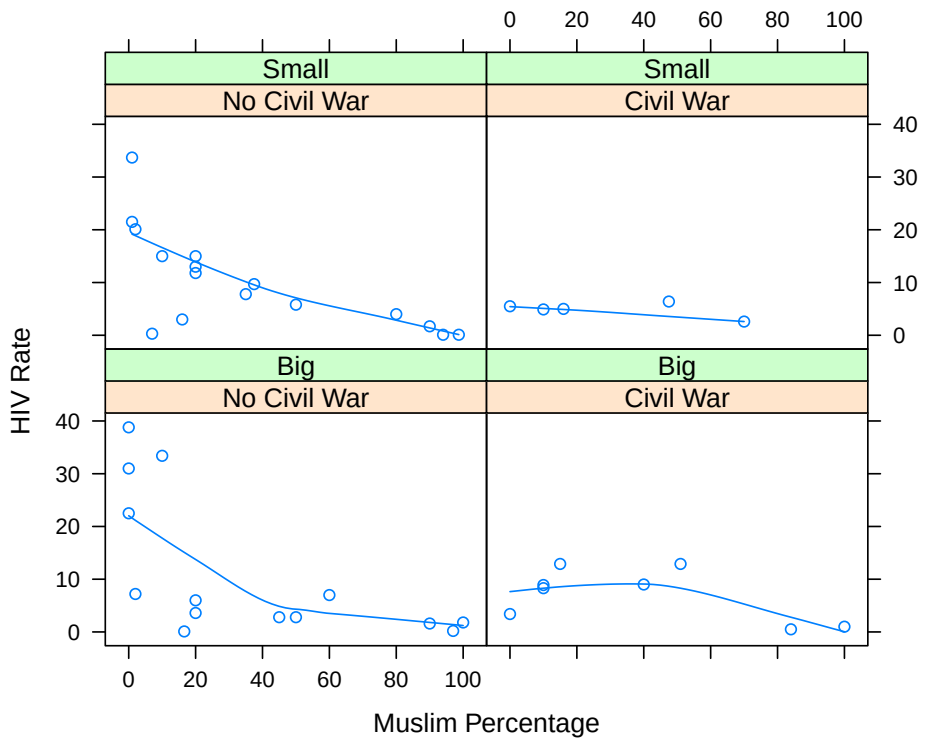


Multivariate Data Display

Once we get to four (and more) variables, we have some hard choices to make. One option is to dichotmize (or, more generally, discretize) one or more of our variables, in order to do conditional plots. If we're willing to do this, it's very easy in R to generate conditional scatterplots using the `lattice` package. For example, if we want to know whether the relationship between Muslim percentages in the population and HIV/AIDS rates is moderated by both the presence of civil wars and country size, we *could* generate two "3-D" scatter plots and compare them. Alternatively, we might divide the countries in our data into "big" and "little" countries (say, by splitting the data at the median) and then generate a four-way scatterplot:

```
> Africa$big<-factor(Africa$population>median(Africa$population),
                     labels=c("Big","Small")) # Splitting population at its median
> Africa$civilwar<-factor(Africa$internalwar,
                    labels=c("No Civil War","Civil War")) # creating a "factor" variable fo
> with(Africa, xyplot(adrate~muslperc | civilwar * big,
              col="black",panel=function(x,y){panel.xyplot(x,y);
              panel.loess(x,y,span=1)},
              xlab="Muslim Percentage",ylab="HIV Rate"))
```

which yields:



Here, we see "little" and "big" countries, as well as countries that do and do not have a `civilwar`. The plot thus suggests that the negative HIV-muslim population relationship seems to hold for small countries, but not for larger ones, and that there are no appreciable differences in the relationships between countries with internal conflicts and those without them.

Other Useful Plots

Finally, note that I haven't even really scratched the surface on the subject of displaying and graphically exploring data. A few topics we didn't even touch on (but that might well be worth looking into) include:

- *sunflower* plots for high-density data (to overcome overplotting),
- `wireframe` plots – 3-D "surface" plots, akin to hybrids of 3-D scatterplots and contour plots,
- *mosaic plots* for visualizing multivariate categorical data, and
- *parallel coordinates plots* – another means of visualizing multivariate (> 3 variables) data, one better for continuous data than mosaic plots.

Your homework exercise (which will be on the github repository by the end of the day) essentially asks you to do a bunch of plotting, describing, and interpreting of data. It is due in class in a week or so.