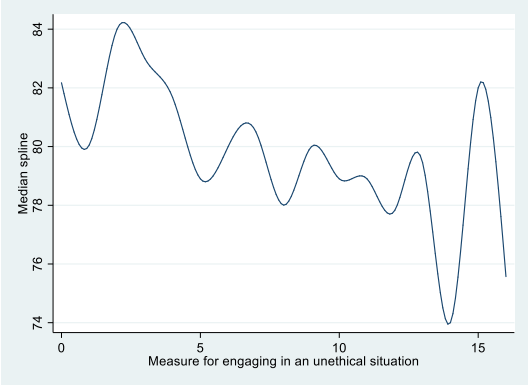
Visualizing data in Stata – Section 6.4
Najib Mozahem

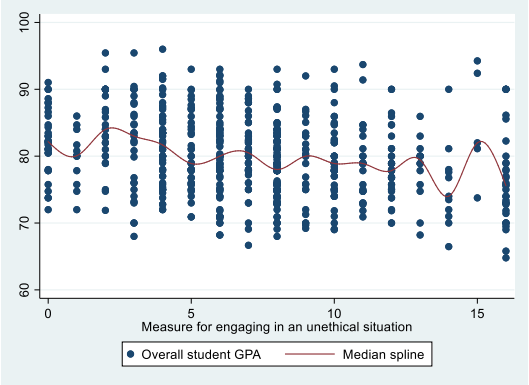
The last lecture ended by noting that scatter plots are not a very good tool when one of the variables is discrete. In such a case, we can instead opt to use a box plot. However, we also saw that in some cases the box plot will not produce output that is clear. This is where splines come in. Stata allows us to draw a curve that fits the data. It is not important to know how the graph is fit, what is important to know is that the command calculates the median value of one of the continuous variable at each value of the discrete variable. As usual, math gets involved, and the output is a graph that is called the median-spline. It is best to see it in action:

twoway mspline gpa engage



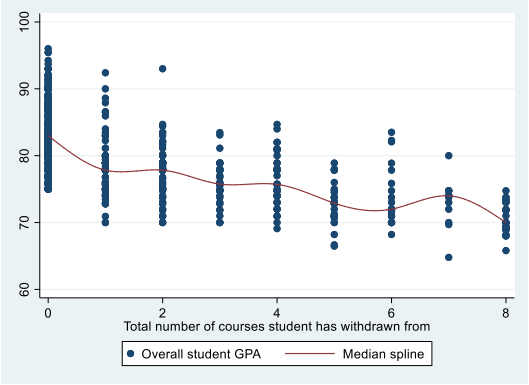
Looking at the graph, we see that there is a tendency for *gpa* to decrease as *engage* increases, but that something strange happens near the end and the dynamic is reversed and then is reversed again. Usually, it is better to draw these graphs together with the scatter plot. As we already know, the **twoway** command allows us to do this:

twoway (scatter gpa engage) (mspline gpa engage)



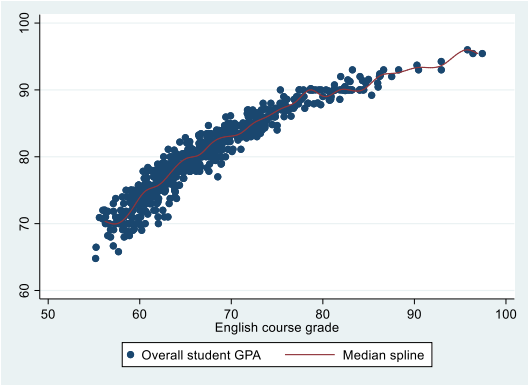
In the last lecture we produced both the scatter plot and the box plot for the variables *gpa* and *withdraw*. Let us now produce the median-spline as well:

twoway (scatter gpa withdraw) (mspline gpa withdraw)

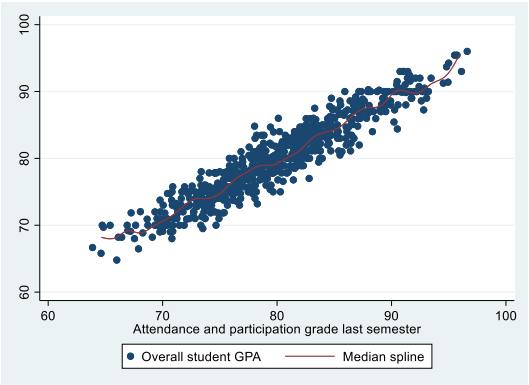


Here we clearly see that *gpa* decreases as *withdraw* increases. Let us now produce the median-spline plots for the three scatter plots that were produced at the start of the last lecture: *gpa* vs. *english*, *gpa* vs. *attendance*, and *gpa* vs. *income*:

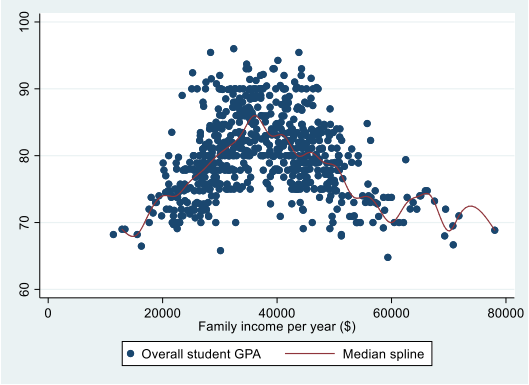
twoway (scatter gpa english) (mspline gpa english)



twoway (scatter gpa attendance) (mspline gpa attendance)



twoway (scatter gpa income) (mspline gpa income)



There is something interesting about the above three graphs, and that is that the median-spline resembles a mathematical function. To be more clear, the relationship between *gpa* and *english*, and the relationship between *gpa* and *income* both seem to resemble a quadratic function. A quadratic function is one in which looks like a parabola. The graph either starts increasing and then starts to decrease, or bend downwards, or the graph starts decreasing and then starts to increase, or bend upwards. In the case of the variable *gpa* and *attendance*, the graph actually looks like a line.

The tools used so far in this section are referred to as nonparametric tools. Nonparametric tools make no assumptions about the data. The data is left to speak for itself. Quantile plots calculate and plot the quantiles without assuming anything about the distribution. Scatter plots on the other hand only plots the values and leaves it up to us to see whether there seems to be a relationship. Splines, or median-splines in our cases, go further than scatter plots because they smooth the data, but the process of smoothing the data is performed without making any assumptions.

However, as we have seen above, it seems that there are some assumptions that we can make in some cases. It seems logical to assume that the relationship between *gpa* and *attendance* is linear, i.e. it follows a line. When we believe that we have sufficient evidence to make assumptions about the data, we can start using another group of tools which are referred to as parametric tools. The next lecture will be about some of these tools.