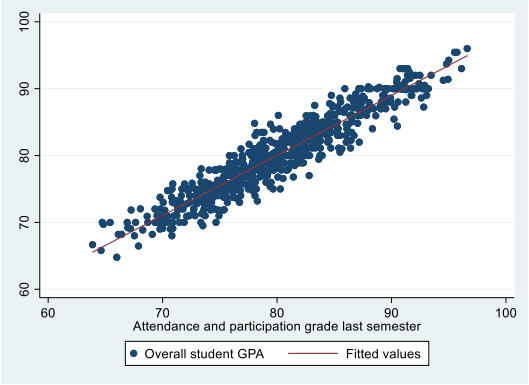
Visualizing data in Stata – Section 6.5
Najib Mozahem

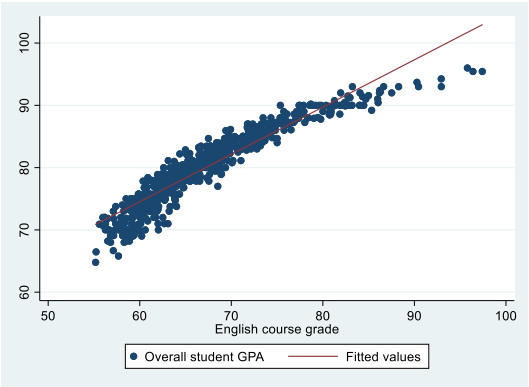
We will start with the simplest case, which is that of a line. Instead of telling Stata to overlay the scatter plot of *gpa* and *attendance* with a spline, we will tell it to draw the line that best fits the data:

twoway (scatter gpa attendance) (lfit gpa attendance)



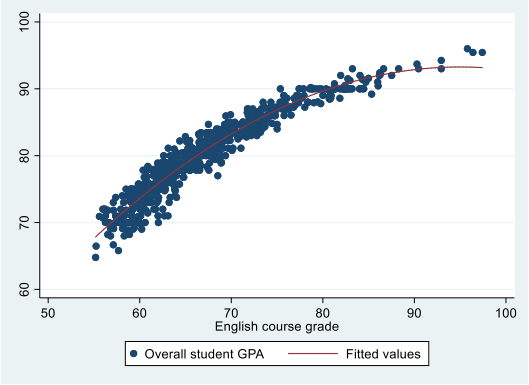
The command used to tell Stata to draw the best fit line is **twoway lfit**. As usual, it is best to overlay this on top of the scatter plot in order to make sure that the line fits the observations. This is actually a very important point because it is not Stata's job to tell you whether you should fit a line or not. If you ask Stata to fit a line, it will draw the best-fit line. However, just because something is the best, it doesn't mean that it is good. As an illustration of this, we already know that the relationship between *gpa* and *english* is not linear. Let us ignore what we know and tell Stata to produce the best-fit line:

twoway (scatter gpa english) (lfit gpa english)



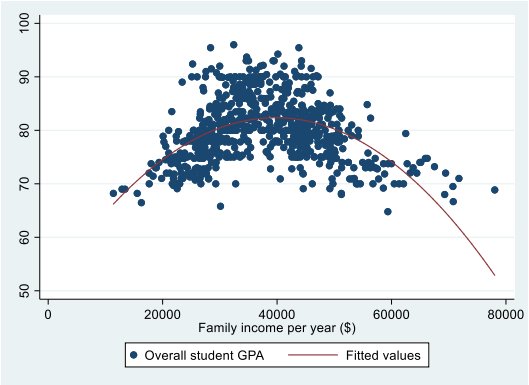
The line does not fit the data well. We see that almost all observations with a grade on English that is greater than 80 lie below the line. One of the criteria for a well-fit line is that the points should lie on both sides and not just one, just as we saw above. Instead of fitting a line, it would make more sense to tell Stata to fit a quadratic function:

twoway (scatter gpa english) (qfit gpa english)



This curve fits the data much better than the line. We can also fit a quadratic curve when it comes to the variables *gpa* and *income*:

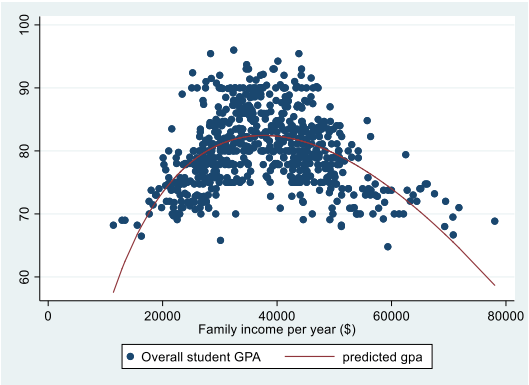
twoway (scatter gpa income) (qfit gpa income)



The curve seems to be doing a good job of fitting the data up until the end, where the observations start to follow on a single side of it.

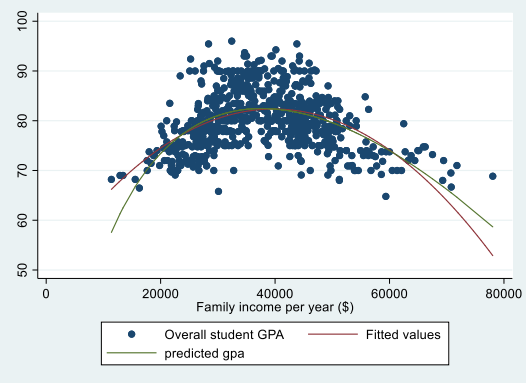
One of the restrictions imposed by a quadratic fit is that the parabola must be symmetric. In other words, the curve must be identical before and after the point at which it changes direction. If you think that this assumption is strong and you would like to relax it, Stata allows you to create a fractional polynomial fit:

twoway (scatter gpa income) (fpfit gpa income)



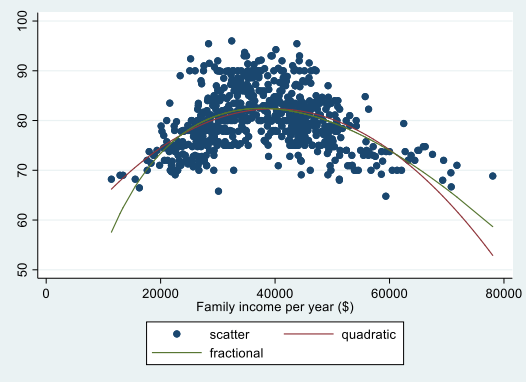
As you can see, the curve is no longer symmetric. It rises much more quickly that it falls. To me, this makes more sense because I believe that the advantages obtained in increases in income in terms of access to resources, have more magnitude than the disadvantages which might result from having too much money, i.e. feeling secure and having other options. We can plot both the quadratic fit and the fractional polynomial fit on top of the scatter plot to help us compare them:

twoway (scatter gpa income) (qfit gpa income) (fpfit gpa income)



The labels used in the legend are not very helpful. We can use the **legend()** option to customize the text (we will cover formatting issues in a separate lecture just like we did for other types of graphs):

twoway (scatter gpa income) (qfit gpa income) (fpfit gpa income), legend(label(1 "scatter") label(2 "quadratic") label(3 "fractional"))



It seems to me that the fractional polynomial fit does a better job at the sides of the scatter plot while they both do a good job in the middle.