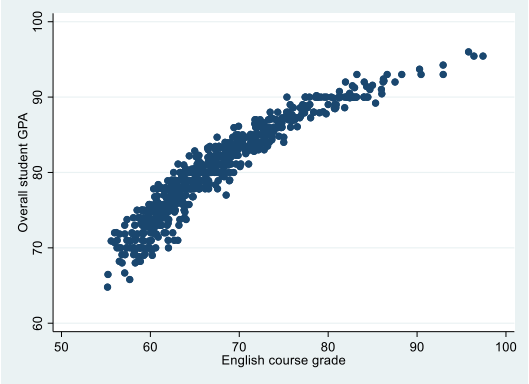
Visualizing data in Stata – Section 6.3
Najib Mozahem

We have now reached the point in the course where we take advantage of Stata's powerful **twoway** command. We have already come across this command in one of the lectures when we wanted to plot the kernel density of a variable across two groups. At that point, it was mentioned that this command allows us to draw different plots on the same graph. This is only one of the powerful options provided by this command.

The **twoway** command allows us to produce many types of plots. The simplest of these is the scatter plot. A scatter plot is simply a graph where the values of two specific variables for each observation are indicated by a symbol. These points tend to be scattered, hence the name.

We start by producing the scatter plot for the two variables *gpa* and *english*:

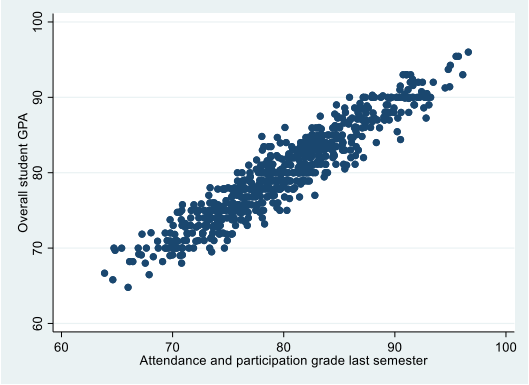
twoway scatter gpa english



Stata has basically plotted the values of all observations for both variables. One of the reasons that scatter plots are powerful is that they allow to us investigate whether there is any relationship between two variables. Looking at the graph, we can conclude that students who get higher grades on English tend to have higher overall GPAs. We also see that the graph tends to start levelling off near the end. What this means is that, after a certain point, having higher grades on the English course does not correspond to large increases in the overall GPA. This makes sense. In order to get good grades in an English teaching institute, you need to have a good grasp of the language, but you don't need to be an expert in it. There is a certain level after which the English grade won't make much of a difference.

Let us now see if attending and participating in classes corresponds with higher GPAs:

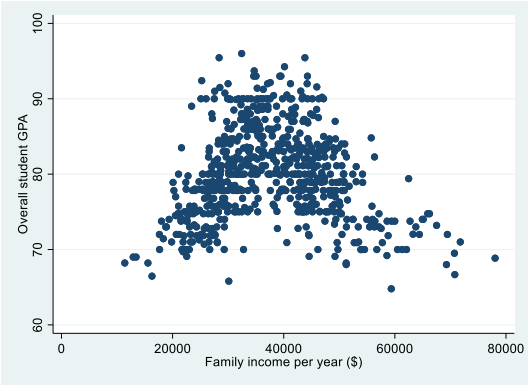
twoway scatter gpa attendance



The graph provides strong support for the notion that attending classes is beneficial to the overall performance on academic studies.

Let us now see whether there is a relationship between the variables *gpa* and *income*:

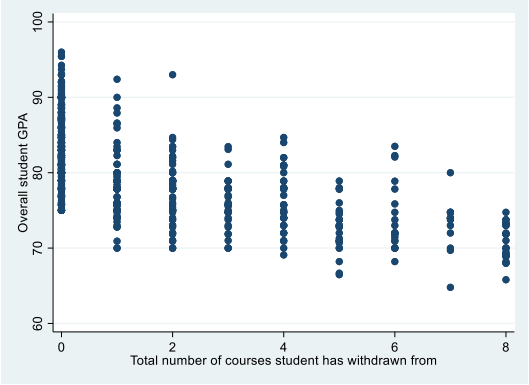
twoway scatter gpa income



The graph looks messier than the ones before it, but there does seem to be a pattern. It seems that students from middle income families tend to do better than students from low income families and students from well-off families.

Let us now see if there is a relation between *gpa* and *withdraw*. Are students with low overall GPAs more likely to withdraw from courses?

twoway scatter gpa withdraw



The scatter plot doesn't seem visually appealing, and there is a very important reason for this, and it is that the variable *withdraw* is discrete. This can be seen by running the following command:

codebook withdraw

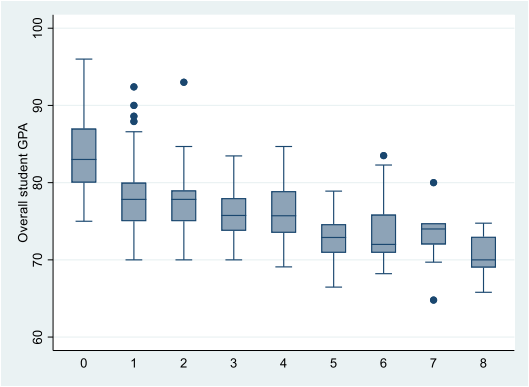
The output tells us that the range of the variable is from 0 to 8, and that there are no decimal values. So the variable can take on one of nine values. Compare this to the output for the variable *gpa*:

codebook gpa

We see that *gpa* has 182 unique values in our dataset. This brings up a very important point. Scatter plots are best used when the variables that are being compared can take on many different values. In the case of the variable *withdraw*, there are many individuals with a value of 0, and each of these has a different GPA. This is why the dots tend to be clustered vertically.

So what do we do in this case? We use one of the tools that we used when comparing variables across groups. Discrete variables are best treated as groups. Therefore, instead of generating a scatter plot of *gpa* and *withdraw*, it would be best to see how *gpa* differs among the different values of *withdraw*. In such cases, I think that it is best to go with a box plot:

graph box gpa, over(withdraw)

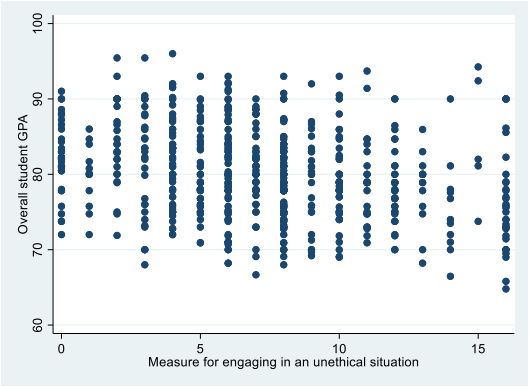


This is a much more useful graph. What we have done is produced a box plot for *gpa* for each of the nine values of *withdraw*, and now we can see an interesting dynamic. The larger the value of

withdraw, the more the box drops down. This means that students who withdraw more tend to have lower GPAs.

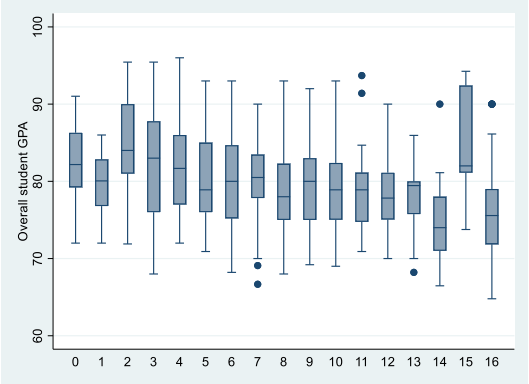
The exact same logic is used when plotting the scatter plot of *gpa* and *engage*. We are interested in seeing if there are differences between students who engage in cheating based on their overall GPA:

twoway scatter gpa engage



Once again we see that the dots are clustering vertically due to the fact that the variable *engage* has only 17 unique values. Therefore, we once again revert to a box plot:

graph box gpa, over(engage)



Do we see a pattern? There is reason to suspect that the value of *gpa* tends to be dropping as the value for *engage* increases, but this is not true for every single box. Unlike in the previous case, it is difficult to reach a decisive conclusion. There is actually a better method in this case, and it is to use a tool that is called a median-spline, which we cover in the next lecture.