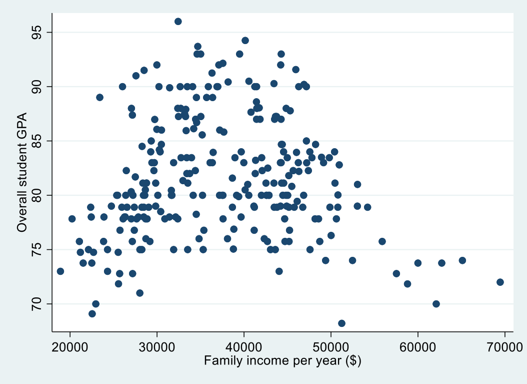
Visualizing data in Stata – Section 7.3
Najib Mozahem

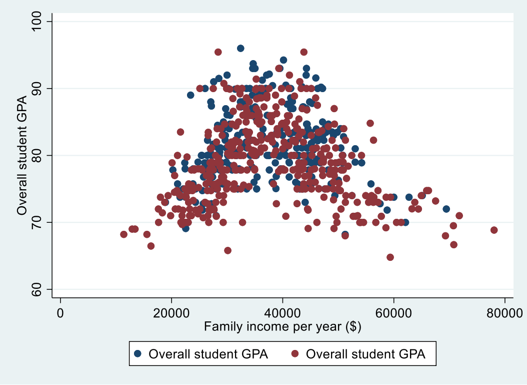
In order to visualize the association between two continuous variables, we use the **twoway** command. This is something that we have already done. What we didn't do was to study group differences in these associations. One of the powerful features of the **twoway** command is that we can use the *if* qualifier with it. The *if* qualifier in Stata is used in order to include only the observations that satisfy a certain condition. For example, if we wanted to graph the values of the variables *gpa* and *income* only for female students, we can execute the following command:

```
twoway scatter gpa income if gender == 1
```



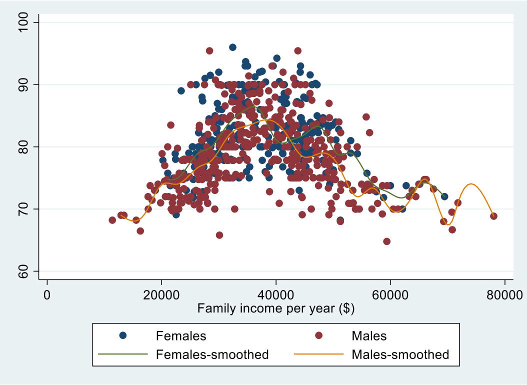
In our dataset, a value of 1 for the variable *gender* means a female. Notice that we use two equal signs when we test for equality in the *if* qualifier. We can also plot the association between the two variables for males on the same graph in order for us to visualize whether there are any differences:

```
twoway (scatter gpa income if gender == 1) (scatter gpa income if gender == 0)
```



Notice that the legend is not really helpful. This is why it would be better to label the legend ourselves. In addition, plotting the splines for each group would allow us to compare the two groups more easily:

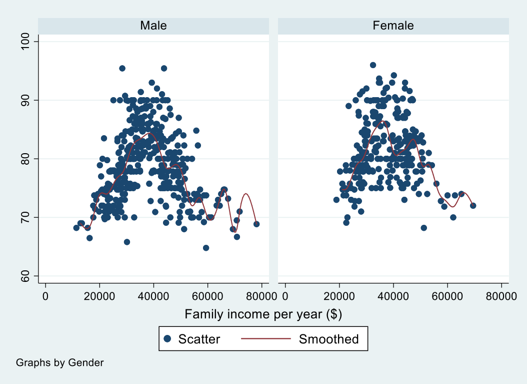
```
twoway (scatter gpa income if gender == 1) (scatter gpa income if gender == 0)
(mspline gpa income if gender == 1) (mspline gpa income if gender == 0),
legend(label(1 "Females") label(2 "Males") label(3 "Females-smoothed") label(4
"Males-smoothed"))
```



In this command we told Stata to draw four graphs. We also used the **legend()** option in order to label the legend in a clear way. Notice that when we specify **label(1 "")** we are referring to the first graph. Since in our command the first plot is for the scatter plot for females, the number 1 refers to that plot. Looking at the smoothed plots of both groups, we see that the curves have a very similar shape, indicating that the same dynamic is observed in both groups, both male and female students from low income families have academic difficulties. The same happens to students from well to do families. The highest GPA grades, in both groups, go to students from the middle of the income curve.

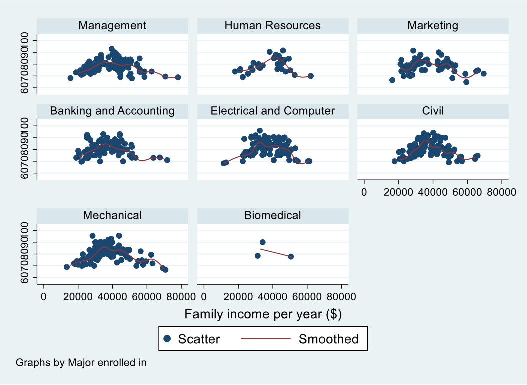
Instead of using the *if* qualifier, we could have used the **by()** option:

twoway (scatter gpa income) (mspline gpa income), by(gender) legend(label(1 "Scatter") label(2 "Smoothed"))



When we use the **by()** option, the legend will only contain two items, since there will be two separate graphs. The choice of which command to use is left to you and to the number of categories. For example, if there are many groups, it would be tidier to use the **by()** option since drawing too many curves on the same graph will be messy. For example:

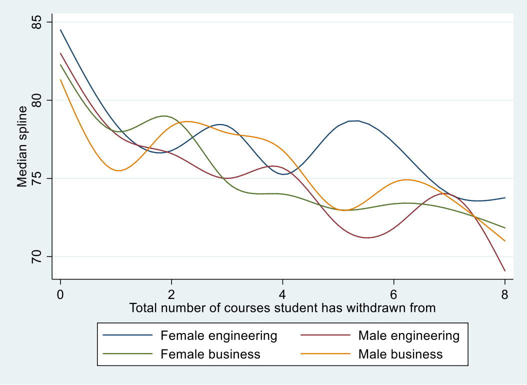
twoway (scatter gpa income) (mspline gpa income), by(major) legend(label(1 "Scatter") label(2 "Smoothed"))



Here we split the graphs using the variable *major*. Since we have eight majors in our dataset, we get eight graphs. Imagine drawing all of these on the same x and y axis. The result would not be attractive.

We can include as many conditions as we want. For example, imagine we wanted to visualize the relationship between the variables *gpa* and *withdraw* while taking into account both the gender of the respondent as well as the college in which he or she is enrolled. This is done using the following command:

twoway (mspline gpa withdraw if gender == 1 & college == 1) (mspline gpa withdraw if gender == 0 & college == 1) (mspline gpa withdraw if gender == 1 & college == 0) (mspline gpa withdraw if gender == 0 & college == 0), legend(label(1 "Female engineering") label(2 "Male engineering") label(3 "Female business") label(4 "Male business"))



In this command, we plotted four splines, one for each category. We have four categories here: female engineering, male engineering, female business, and male business. We see that in all cases, the students who withdraw the most tend to have the lowest GPAs.