Education, Union Formation and Childbearing: Descriptive Diagnostic Plots

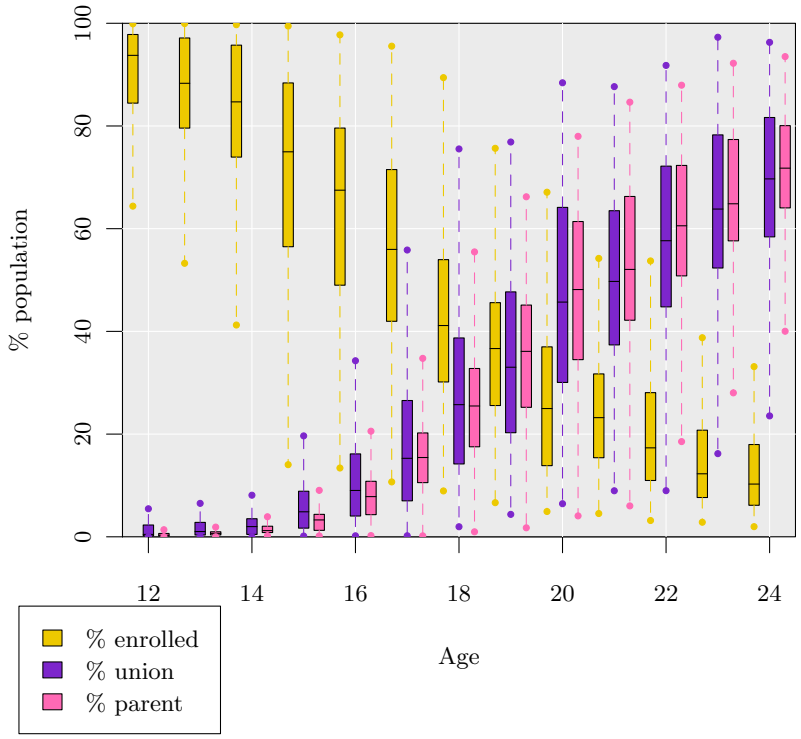
Tim Riffe

November 7, 2011

Abstract

This document is intended to serve as a graphical overview of WORLDFAM data presently being used to describe the relationship between school attendance (enrollment) and the two reproductive transitions of union formation and childbearing. This analysis is crossnational, and includes data from many more populations than all prior studies on similar topics. This has been possible by the combining of two sources, IPUMS and DHS data, both from years close to the year 2000. In order to include this large set of countries it has been necessary to limit our study questions to those that can be answered with current status information. Rather than to establish cause and effect, or propose new mechanisms that might explain the patterns present in these data, we will be content to take a first glance at the patterns themselves. The challenge will be to find ways to graphically display large amounts of data while being able to visually separate different dimensions of the data, such as age and individual country datapoints.

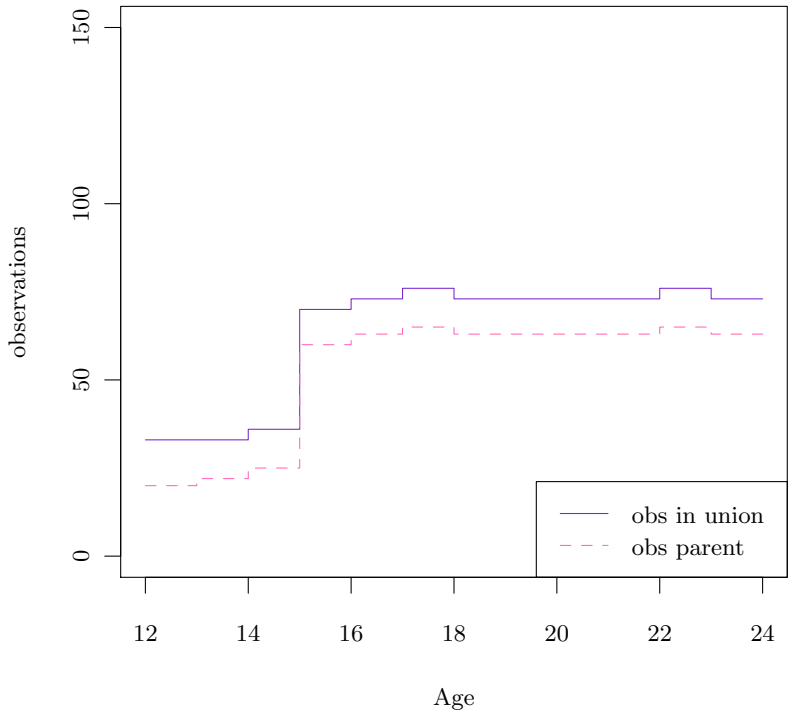
Contents



1 Figure 1 boxplots

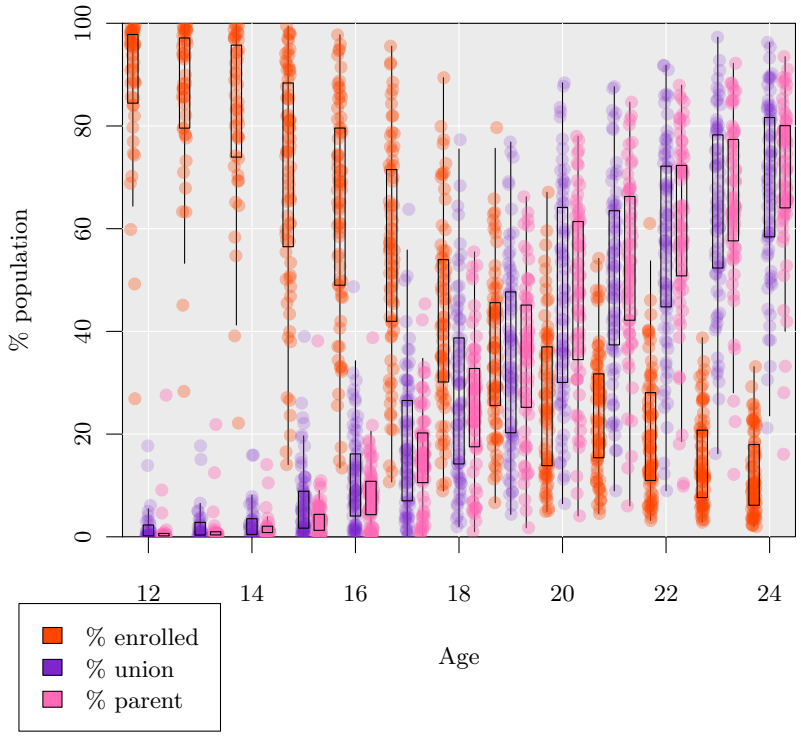
1.1 Figure 1, Females, % Enrolled vs % in union, % parent

We start with boxplots, about as aggregate as you can get. Strictly speaking, the different box colors cannot be directly compared, even within the same age, because there are different observations counts for each age/variable. All valid observations were included to produce this plot. See the following figure for an idea of how observation availability may be influencing the present plot. Recall, these boxes are rough overviews of distributions over countries and nothing more, so even if the same countries were in each age and variable, it is still a long step to infer a standard curve, or age pattern, behind these distribution summaries. To be clear, the central line in each box is the median, the upper boundary the 75th percentile, and the lower boundary the 25th percentile, the maxima are the *smaller* of 1) the maximum value or 2) the 75th percentile+ 1.5* the interquartile range (IQR). This is standard practice for boxplots, since the point is to reduce outlier noise. Further onward we will see the noisy version of this exact same plot. Review this plot taking into account the following plot.



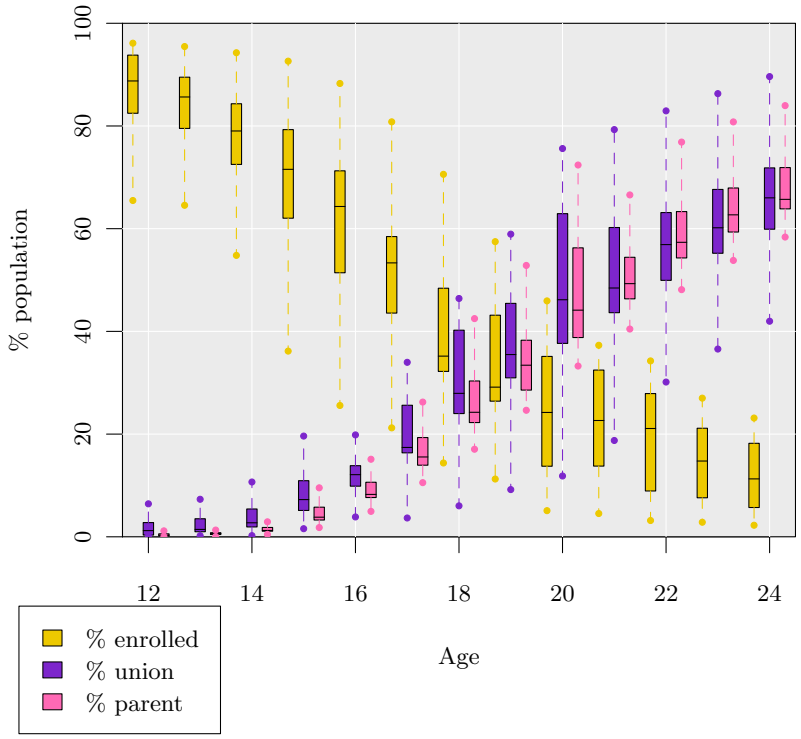
1.1.1 Figure 1.1 female boxplots case counts

This plot displays the case count used to cacluate the quantiles in Figure 1.1. First, the observations available for females, year 2000 round are different in most ages between the two variables "in union" and "has child". Second for ages 12-14 we have far far fewer country observations than at higher ages. The later is worth mentioning if when presenting this in person, and the former we may wish to remedy by limiting ourselves to those countries for which we have both variables. This could reduce the observations within each age to below 80 at ages 15+.



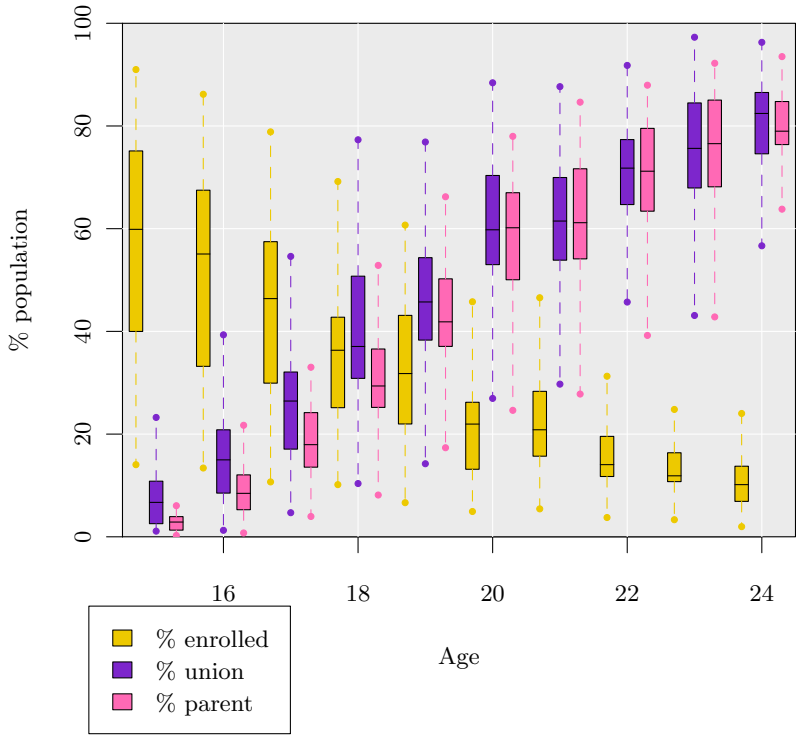
1.1.2 Figure 1.1 female boxplots, all points

For the sake of thoroughness, have a look at what I call the noise-plot behind Figure 1.1. The case counts from Figure 1.1.1 are here still relevant. Note the different color for enrollment, since yellow doesn't show through very well with transparency. This plot should make clear 1) that quantiles are not indicative of natural breaks!, 2) that there is massive dispersion, 3) that most of the time the extreme points displayed in Figure 1.1 coincide with the maxima and minima of the observations- this is less the case with the younger ages for which we have fewer observations. In the later ages where you don't see the black lines among the minima, this is because the transparent points are so heavily stacked on top of the lines that it covers the line up. That's not bad- if you can't see the line because its covered with points then that means the points fall within the line.



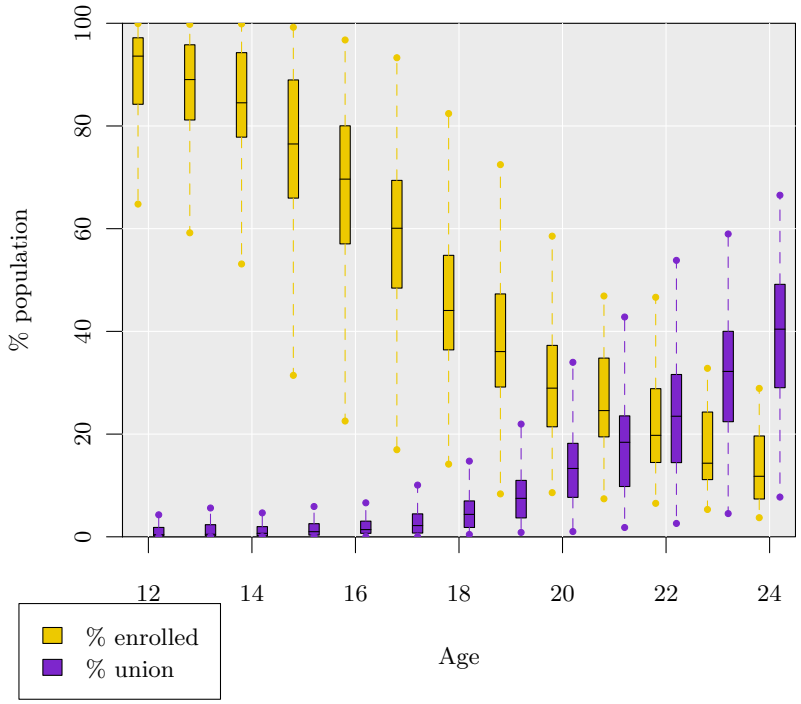
1.1.3 Figure 1.1c female boxplots, only countries with info at all ages

Figure 1.1.2 repeats Figure 1.1, including only those samples with non-missing information for union status and parenthood for all single ages from 12-24. This drastically reduces the observation count to 18 within the figure but removes any observation heterogeneity present between ages and variables. That is to say, the same 18 samples appear in each age and have non-missing values.



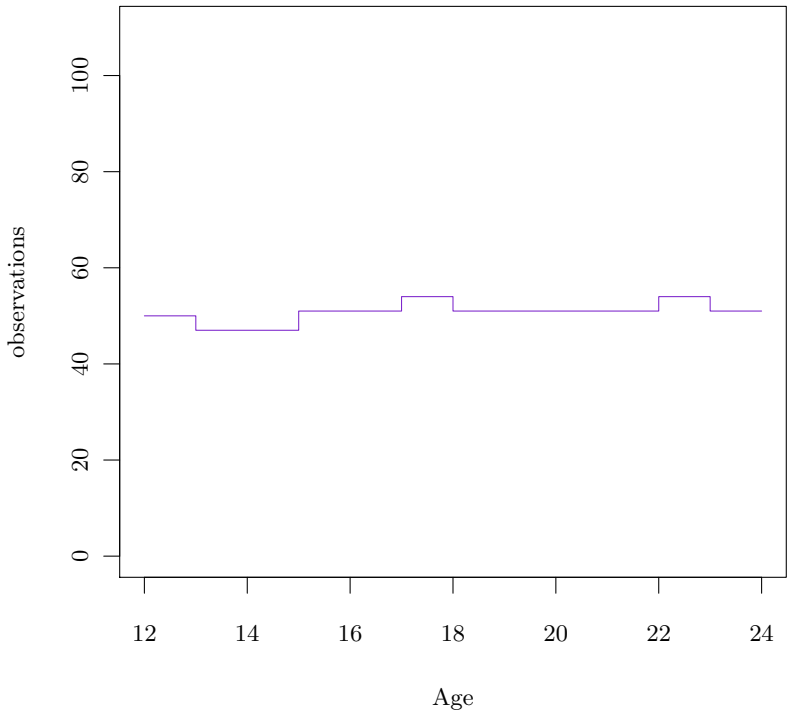
1.1.4 Figure 1.1d female boxplots, only countries with info at all ages >= 15

Figure 1.1.3 repeats Figure 1.1, including only those samples with non-missing information for union status and parenthood for all single ages from 15-24, increasing the number of samples from Figure 1.1.2 from 18 to 27, and is also free of distortions that might be due to irregular inclusion of observations. If we were to increase the first age to 16 or 17, we would gain a few more samples but lose information.



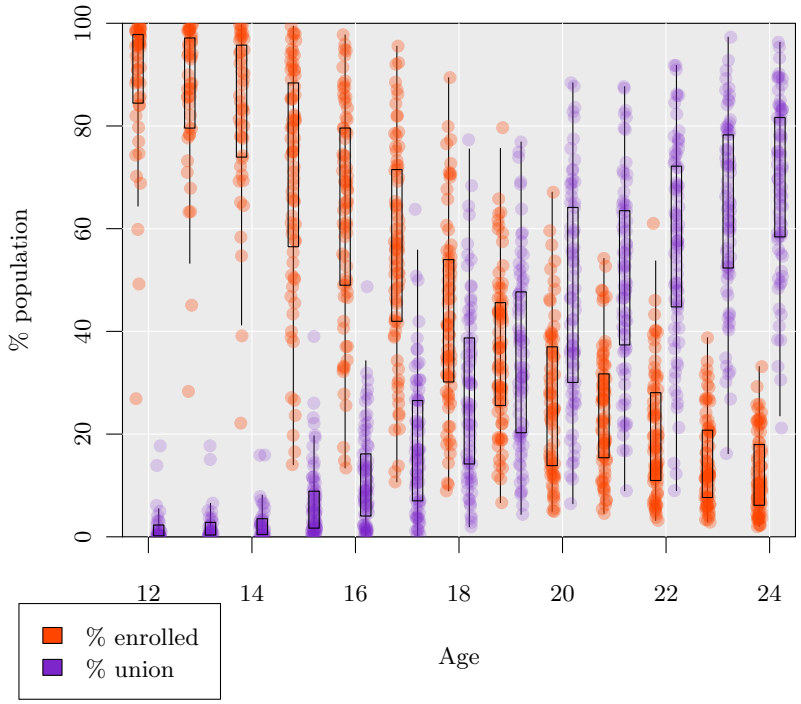
1.2 Figure 1, Males, % Enrolled vs % in union

Now, we repeat the same exercise for males, producing the same three plots from above. Note that males have no information for childbearing, so the figure may appear somewhat more sparse. This plot is of course no solid proof, but nudged, winks and points in the direction of there being much stronger separation of roles for males than females. This may be due to 1) the male marriage curve starting later in life *anyway* or 2) the male marriage curve being *dependant on* males' human capital (or something like that), thereby strengthening the role incompatibility hypothesis, or 3) codetermination, or none of these... This will all become clearer when we look at true bivariate relationships, and when we take a second look at these boxplots, splitting on school attendance.



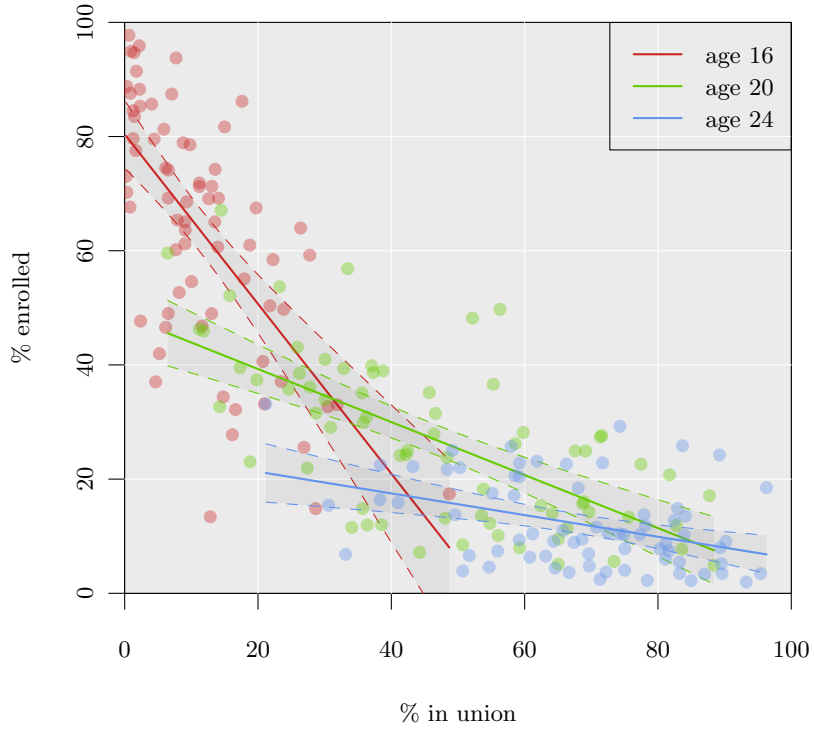
1.2.1 Figure 1.2 male boxplot case counts

Here, a glimpse at the case observation counts used in the above plot. This is less interesting than for the case of females because we can't compare it with fatherhood. The same scales are used. Note that in general we have less information on males than on females, but *more* country-observations at younger ages. Based on this, Joan thinks that this is perhaps a DHS data problem when combining files, and he is presently looking into regenerating the data behind these figures. This is no big deal for the present work being done, since everything can be regenerated with a click, and later versions of this very document evolve in sync with newer versions of data.



1.2.2 Figure 1.2 male boxplot, all points

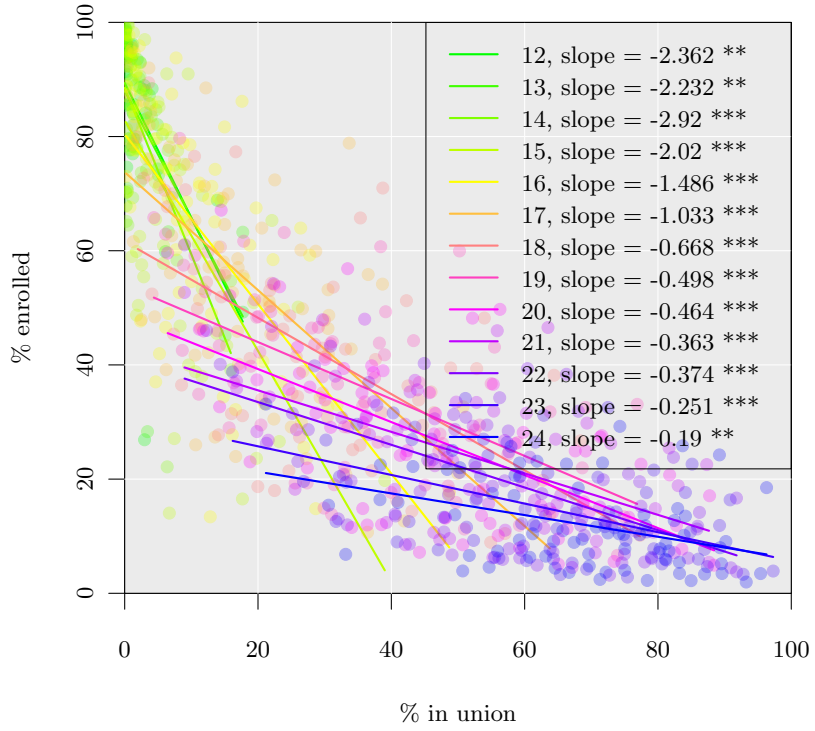
Again for the sake of thoroughness, let's have a look at the noise plot behind the male box plot. The only thing possibly shocking about this are the two observations at ages 12 and 13 that are above 15% married already (Rwanda and Senegal)- it's possible, but Joan is presently looking into the case counts from which these figures were derived.



2 Figure 2 Scatterplots

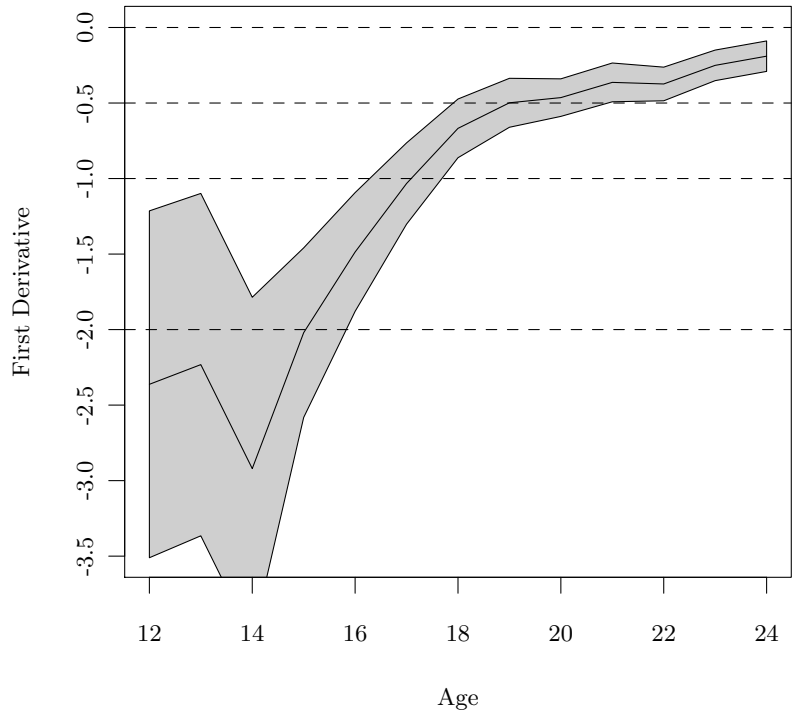
2.1 Figure 2.1, Females, Enrollment vs in Union

Starting always with females, we first look at the bivariate relationship between the aggregate proportion of a population enrolled in school versus the proportion in union. Three representative ages, indicated by color, have been selected to include in this plot in order to reduce clutter, with each point being a country. Thus, each country is included a maximum of three times in this plot. We wish to see whether there is a clear (significant) relationship between the proportion enrolled and the proportion in union, and how this changes with age and so for each of the 3 clouds of points and OLS line is fit, and 95% confidence bands are drawn- each slope is significant. The present plot includes a total of 93 country-observations (some countries may appear more than once, if say, they have a 1998 and 2001 sample, or some such thing). The axes have been chosen as such because it seems logical that school attendance slides downward with age, but really the plot would be identical but transposed were we to flip the axes.



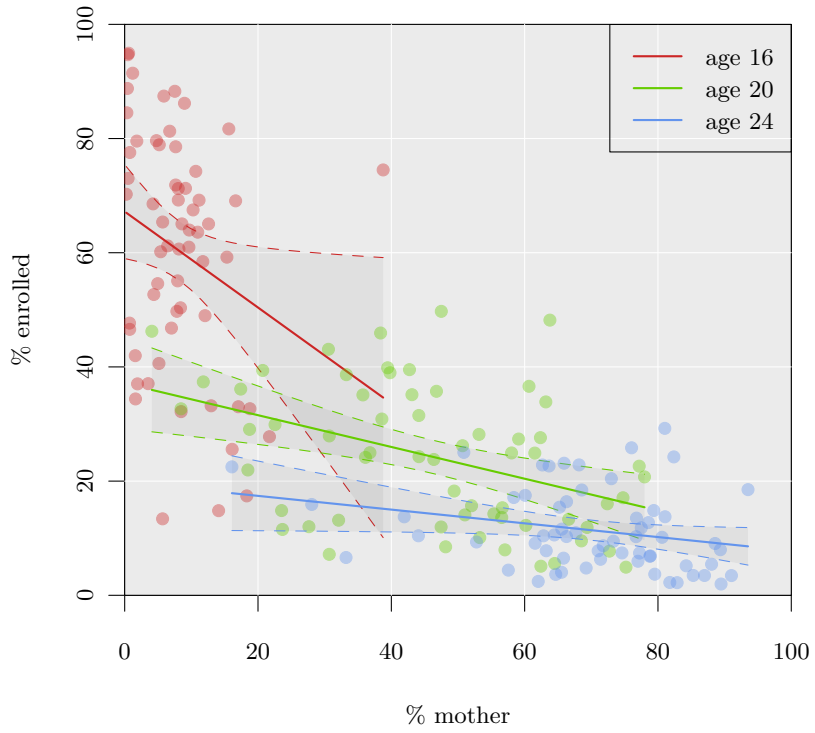
2.1.1 Figure 2, Females, % Enrolled vs % in union, single ages

Now we redraw the exact same plot, same variables and same data sources, but including each age. Within each age, and OLS line is drawn, but no confidence bands are drawn. The slope starts off as steeply negative holding steady until around age 15, then drops rapidy until around age 19. The change in slope over age will be examined in a later figure. Color is used to differentiate between ages, with greenish hues indicating young ages (13ish), yellows and reds for the middle ages (14-19), and blues and purples for the upper ages in our range (early 20s). Transparency is used to reduce noise in the clouds of points.



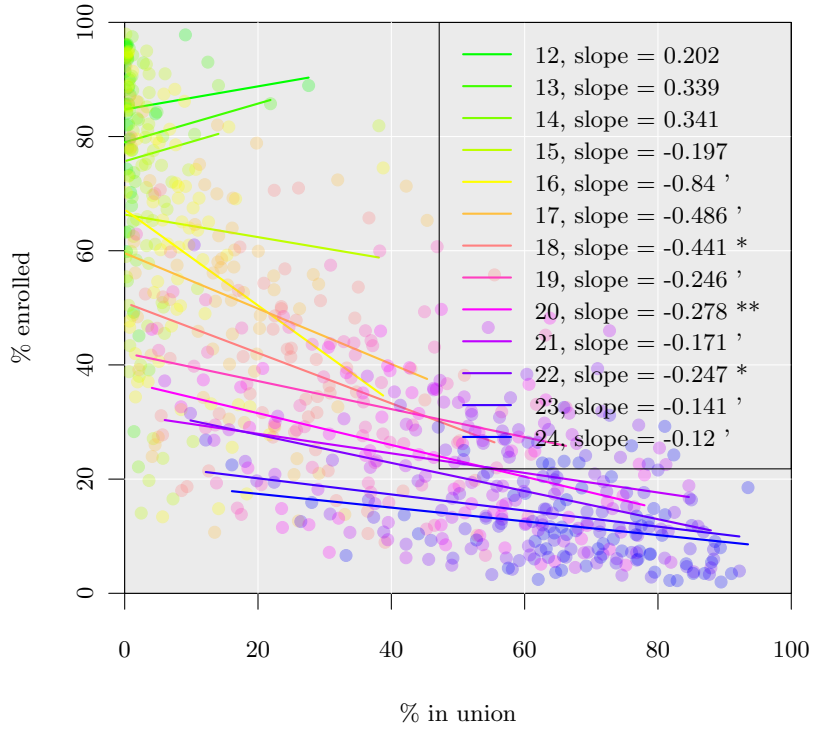
2.1.2 Figure 2, Females, % Enrolled vs % in union, change in slope of age-specific OLS

Questions were raised about the age pattern of the *slope* of the above age-specific OLS lines, displayed in Figure 2.1.2. This figure, although uncommon, has a straighforward interpretation: Its steepest segments are the ages where change accelerates the fastest (the second derivative). In the first place, these are from ages 14-16, followed in importance by ages 16-19. Eyeball compare this with colorful Figure 2.1.1 and this becomes apparent. Another useful reference is the point at which this curve crosses -1 on the y-axis. This age, somewhere between 16 and 17, is the age at which 10% higher average enrollment predicts 10% lower probability of being in union- *across* countries and *within* that age range. Where the line crosses -2, enrollment is to be understood as being very predictive: 10% higher enrollment in a particular country predicts 20% lower probability of finding oneself in union: this is at age 15, but also conceivably the case (within confidence bands) at earlier ages as well. By around age 18, the strength of relationship has slacked off to -.5, meaning that 10% higher enrollment predicts a meer 5% lower chance of being in union, and this relationship remains negative and significantly different from zero until the highest age considered in our study, age 24.



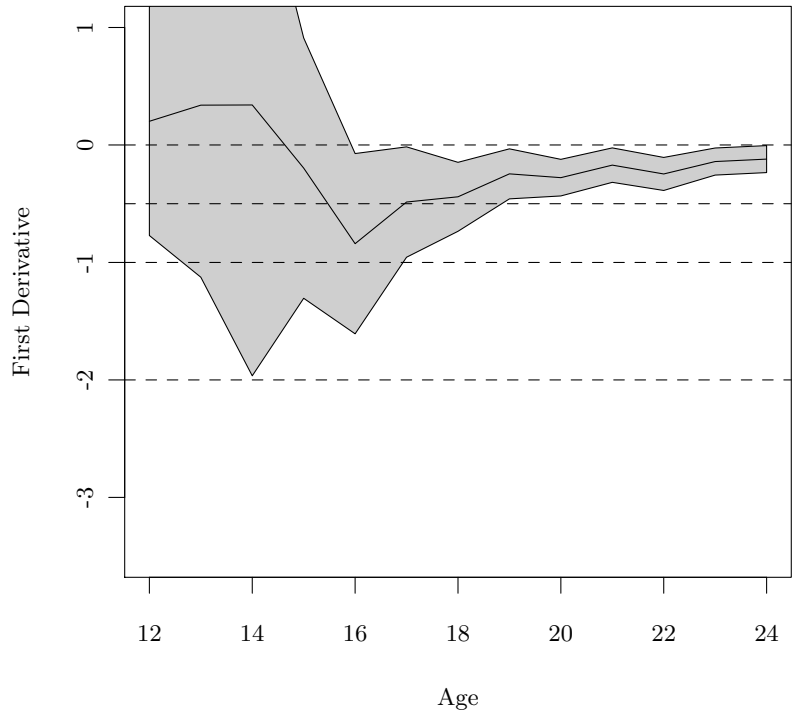
2.2 Figure 2.2, Females, Enrollment vs With Child

We repeat the above three plots again for females that have had a child. The country observation count indicates that a country was present in at least one of the age-cuts. All three OLS lines are significant at the 95% level.



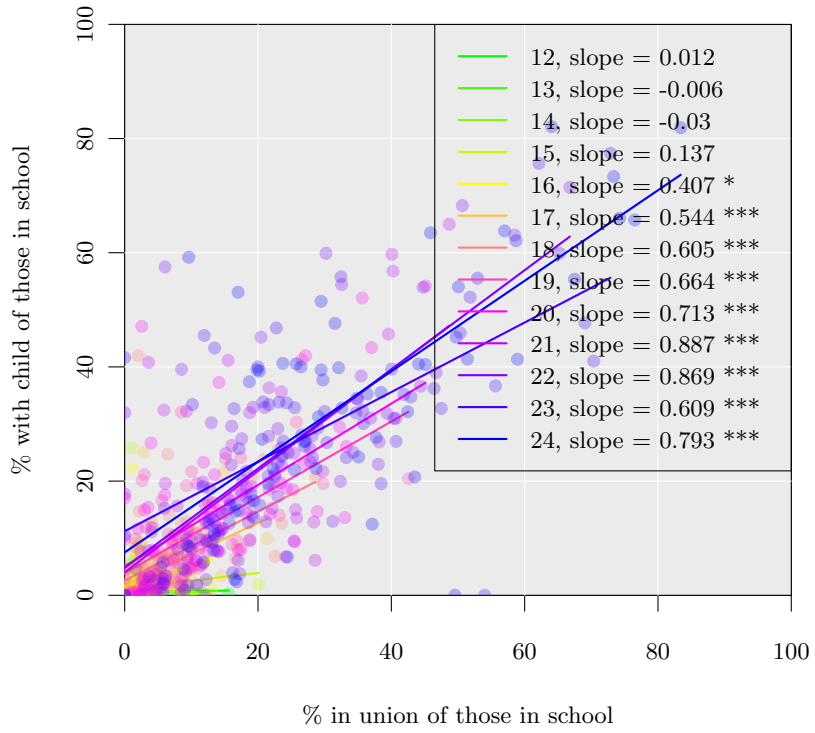
2.2.1 Figure 2, Females, % Enrolled vs % with Child, single ages

Likewise, why not fit an OLS line to each age separately, checking slope and significance. Is there at age at which the relationship between school enrollment and having had a child becomes or ceases to be significant? Yes: age 16. At ages 14 and under we see no significant relationship (in fact the slopes of those lines (greens) can be ignored altogether)- this is prior to and around the age of menarche, and there are likely very few cases in any dataset that would shed light on those ages. The single star at age 16 indicates a mere 90 % significance, while all later ages are *very significant*. Among the significant ages, the pattern is similar to that for union formation. In following, we plot the age pattern of the slopes of these colored lines.



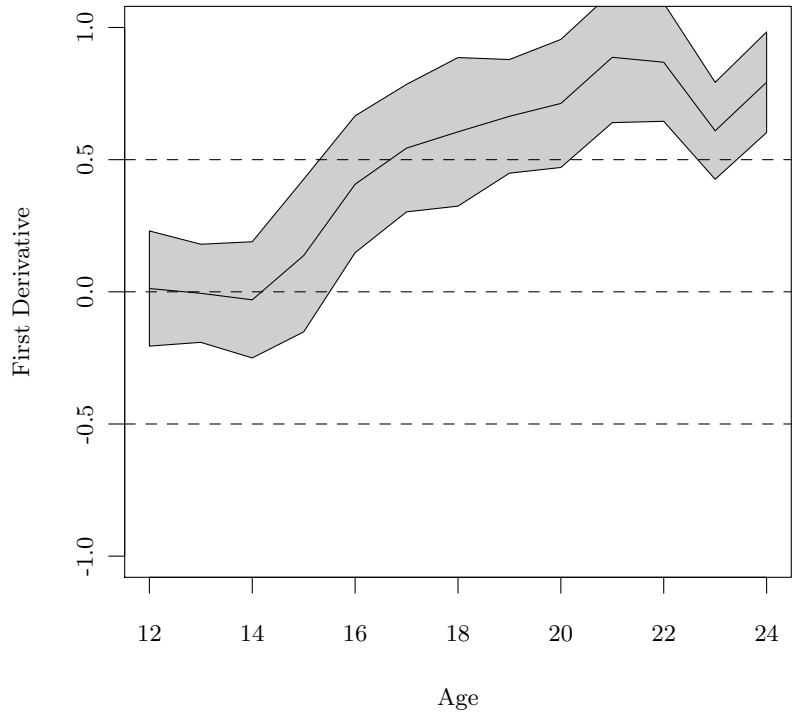
2.2.2 Figure 2, Females, % Enrolled vs % has child, change in slope of age-specific OLS

Figure 2.2.3 summarizes Figure 2.2.1 nicely: nothing is significant until age 16. There is no dramatic weakening of relationship, as was the case with union formation. From age 18 onward, the strength of relationship is very similar to that of union formation, i.e. 10% more enrollment predicts 5% (or less) lower probability of having had a child. We do not worry about the very young ages, as that absence of significance is above all, physiologically determined. Culture can only stack on top of that, but even if childbearing were outright encouraged by some culture at age 13 or 14, we still wouldn't see a significant relationship, and we therefore have no evidence for or against the effects of school attendance in those ages. However, we know that being in union increases the odds of and generally preceeds childbearing, and the relationship between enrollment and union formation *is* in the right direction and very significant in the young ages.



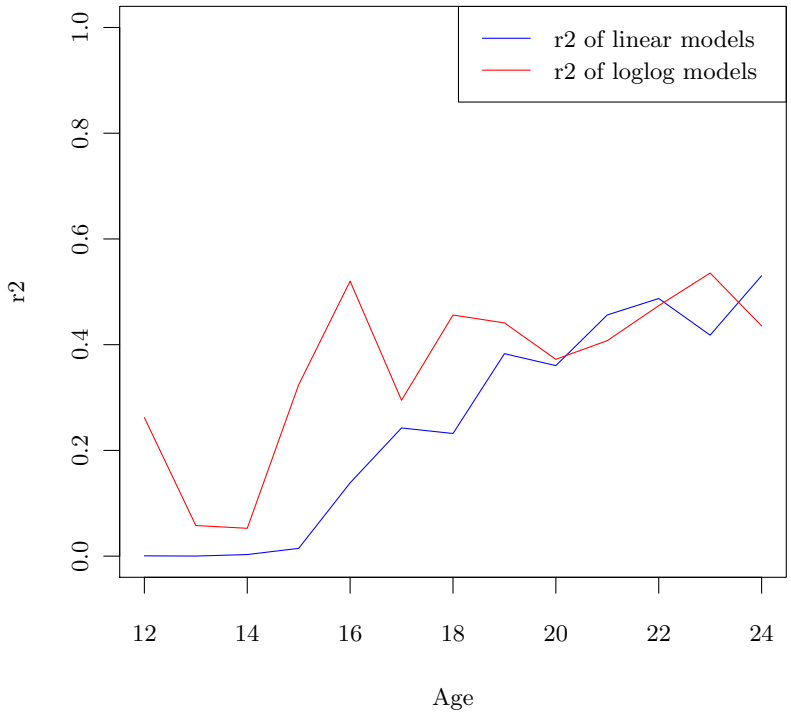
2.2.3 Figure 2, Females % with child vs % in union, both for those in school, single ages

As requested, here is a scatterplot looking only at those in school: % in union vs % with child. There are indeed various samples showing higher proportions with children in school than in union in school. Is this a finding or a trick?! Another observation: this plot is not very informative in general. It turns out the the linearity of the relationship between these 2 variables is much stronger when both are logged. See following Figure.



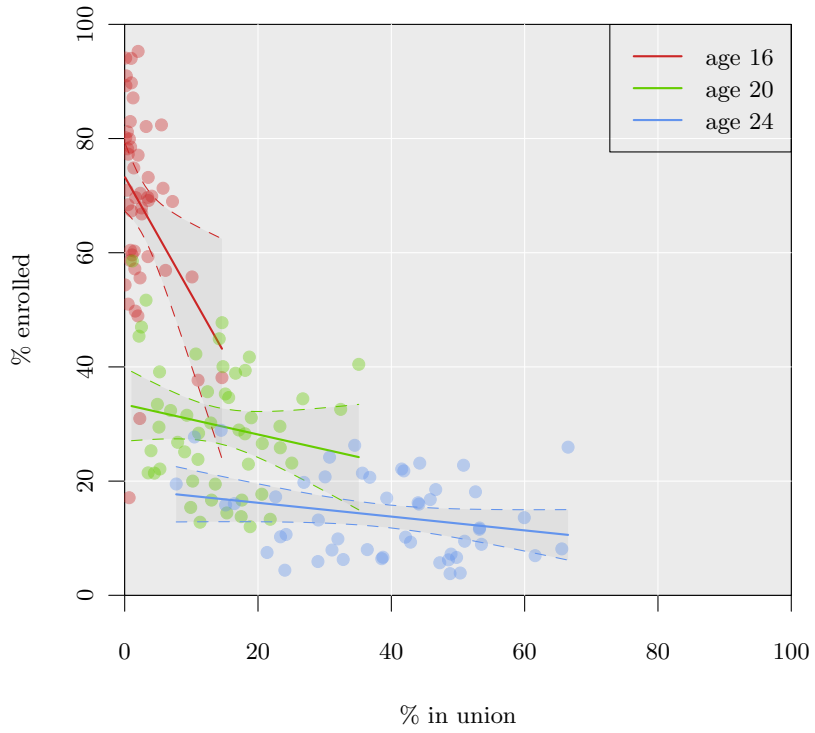
2.2.4 Figure 2, Females % with child vs % in union, both for those in school, change in slope of age-specific OLS

The change in slope of the OLS lines derived from Figure 2.2.3 data. Starting at age 16, the slope is significantly positive and gradually increases.



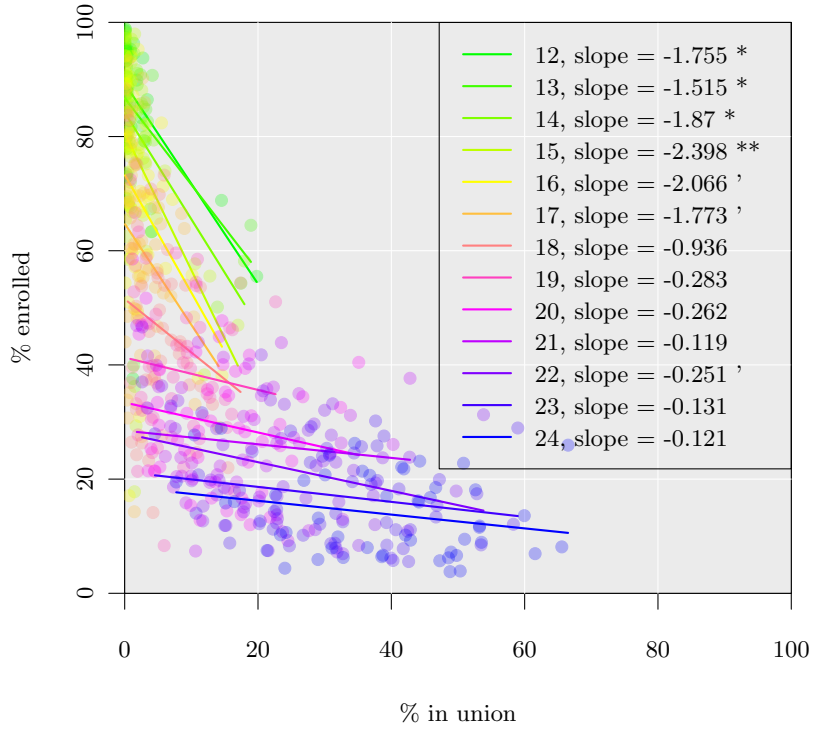
2.2.5 Figure 2, Females % with child vs % in union, both for those in school, comparing r^2 over age between linear and log-log models.

Log-log models are of course more difficult to wrap your mind around, but I figured it worth sharing this observation: If we take the r^2 value for each OLS line from Figure 2.2.3, it is lower at each age than the corresponding r^2 of OLS lines fit to the same data after logging. The figure of the logged data is not shown, but in general more ages turn out significant (slopes and CI not shown, but you get the idea). This may also turn out to be the case with other plots where one or both variables are crammed into a corner in certain ages, and I can explore this further if you wish.



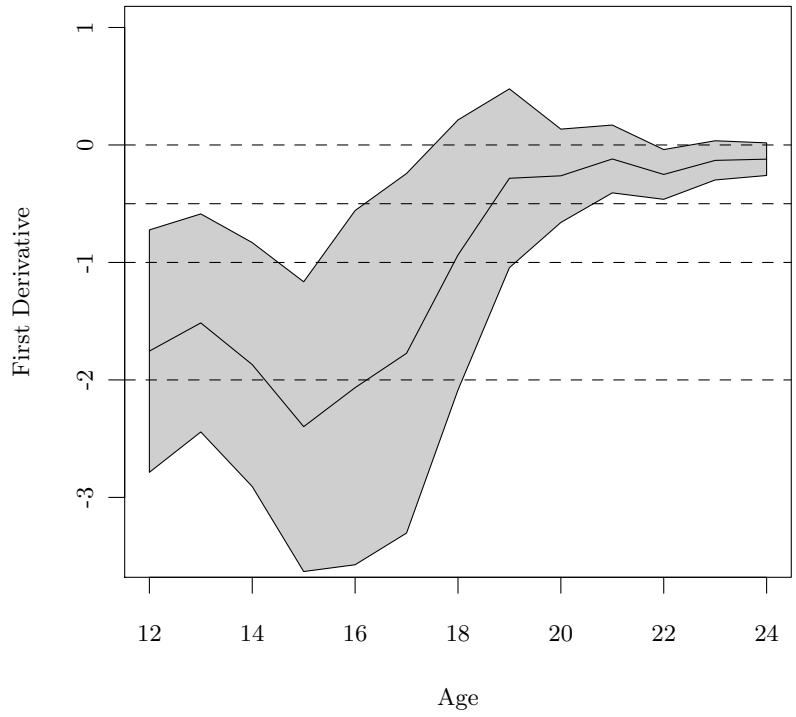
2.3 Figure 2, Males, Enrollment vs In Union

Figure 2.3 repeats the bivariate exercise for males enrolled versus in union, we see that age 16 shows a very significant relationship, while ages 20 and 24 show no relationship. The age-specific OLS to follow will provide more information about what's going on.



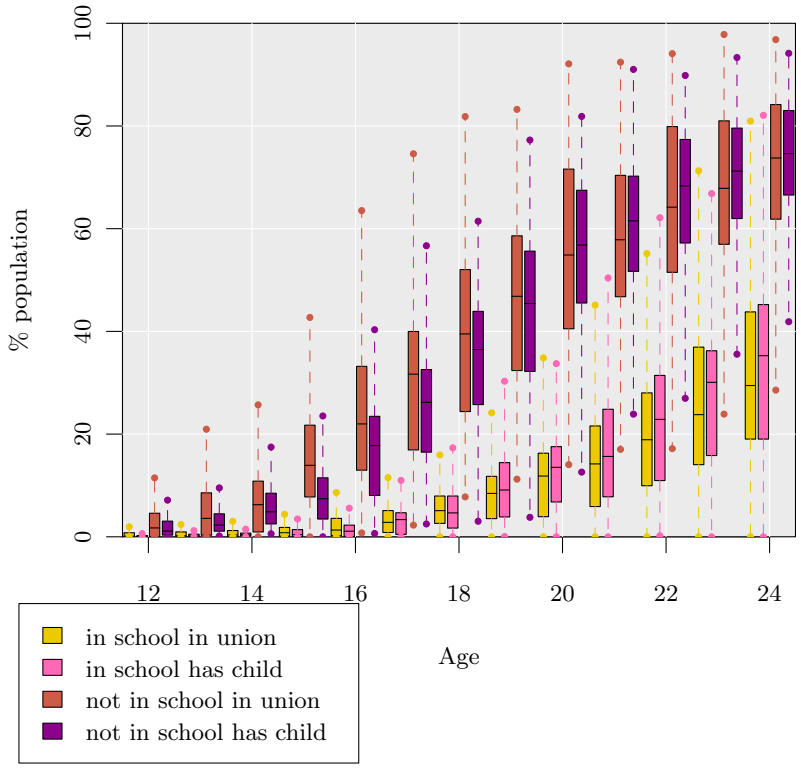
2.3.1 Figure 2, Males, % Enrolled vs % in union, singles ages

This is where the story is at: For males there is a significant negative relationship between aggregate school enrollment and percentage of the population in union *until* age 17. The greens and yellows change their intercept, but appear rather parallel. The following derivative plot will probably tell us what's going on. In short, we can *suspect* that role incompatibility is strong among young males (perhaps under rule of shotgun?), and rather non-existent from the late-teens and onward. Later, we'll disaggregate by enrollment and see whether this is just a heterogeneity trick.



2.3.2 Figure 2, Males, % Enrolled vs % in union, change in slope of age-specific OLS

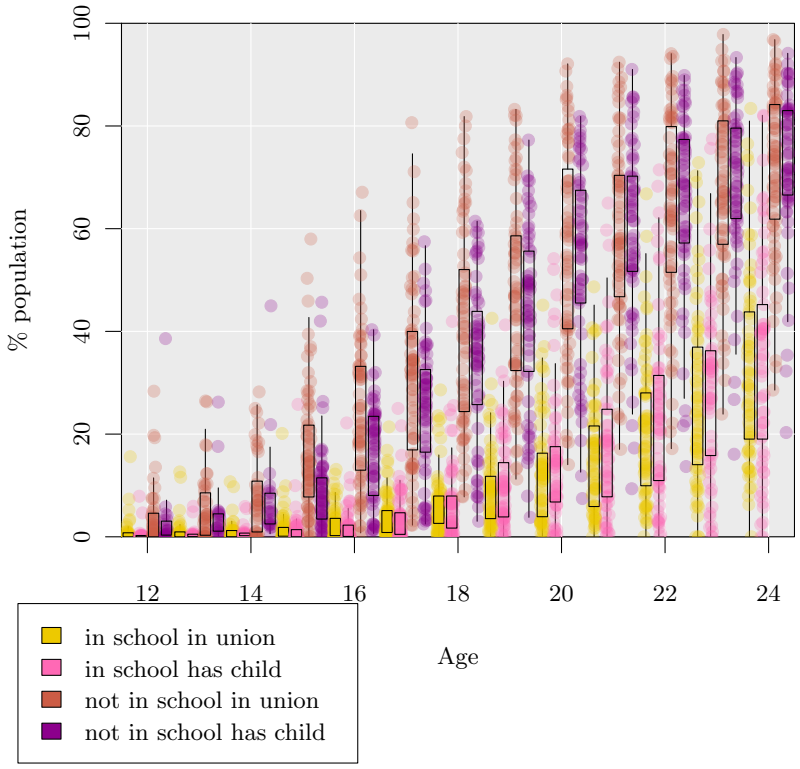
Note first that the confidence bands are all-around wider for males than they were for females. This is partly because we have 20-30 fewer observations available for males than for females and partly due to the relationship being less regular across the set of sampled countries, possibly due to great differences between populations in typical male lifecourses. Shall we say, until age 16 a country's having 10% higher school enrollment predicts around 20% lower levels of in union among males. Thereafter, school enrollment is simply not a good predictor. This does not worry us- in many marriage regimes, males are the primary earners while married. Differences is individual time alotment might not be all that great between *working* and being in school. Why then shoul dbeing in school preclude marriage? One might say that working predicts marriage because males need income to be marriageable, but then if education predicts income, then we would expect smart ladies to snatch up males in school (or have it arranged so!): hence no role incompatibility later on.



3 Figure 3 boxplots

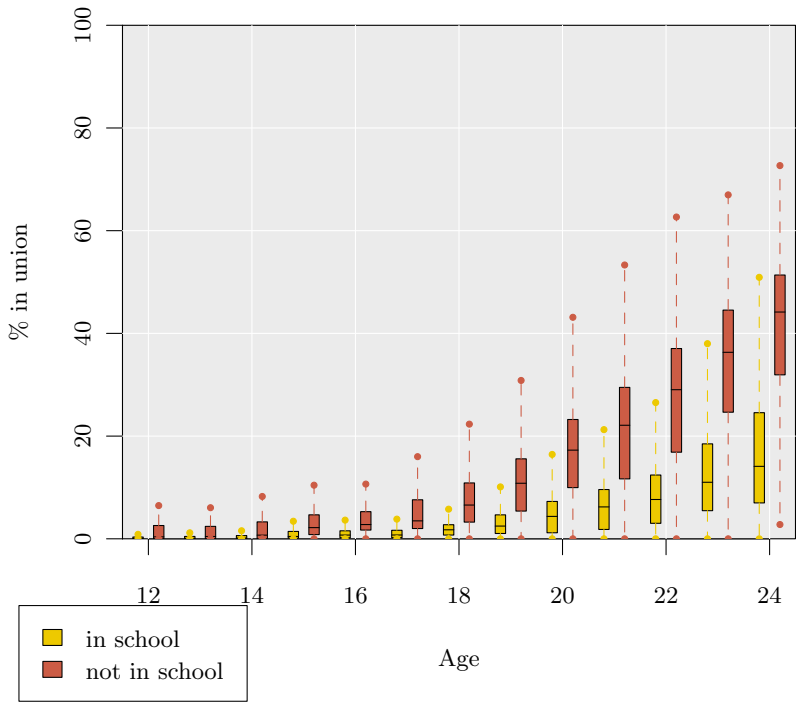
3.1 Figure 3, Female With Child and In Union by Enrollment, single ages

All of the above plots are based on aggregate data, and so beg the question as to how much these patterns hold when looking within groups: how much of the aggregate pattern is being determined by compositional effects? We first divide the data into two groups- those attending and those not attending and look again at proportions in union and that have children, females then males. Figure 3.1 makes obvious the separation between those in school and out of school in terms of the timing of union formation and childbearing. This plot was based on the same observations as previous boxplots, and this explains the irregularity between the in union and childbearing bars. "In union" should always be a bit higher than "has child", both in and out of school. Joan has verified that this is the case when looking only at the countries that have information for both variables. I will not redo the observation counts plot, but do include a noise plot for reference in Figure 3.1.1.



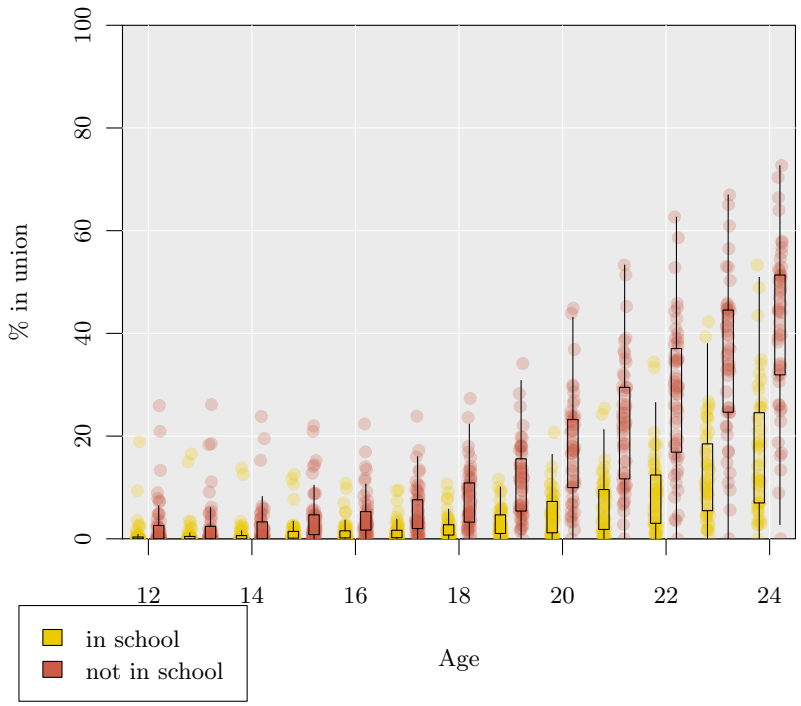
3.1.1 Figure 3b, Female With Child and In Union by Enrollment, single ages, all observations

Figure 3.1.1 plots the same data as the prior plot, but includes points for all country observations. In this case there are certainly more observations falling outside of the stems. This might be explicable by sample size: it will be worth comparing this with the case count tables. Specifically, the two high points for girls in school and with children at age 12 and 13 are Iran, with more than 5% each. Out of school and with child at age 12 is also Iran, with almost 40%, and Iran and Thailand at age 13. I'm now thinking that for those countries where we include various samples for the year 2000 round, e.g. 1998 and 2001 DHS surveys, we ought to think of pooling these into a single 2000 round sample for our purposes.



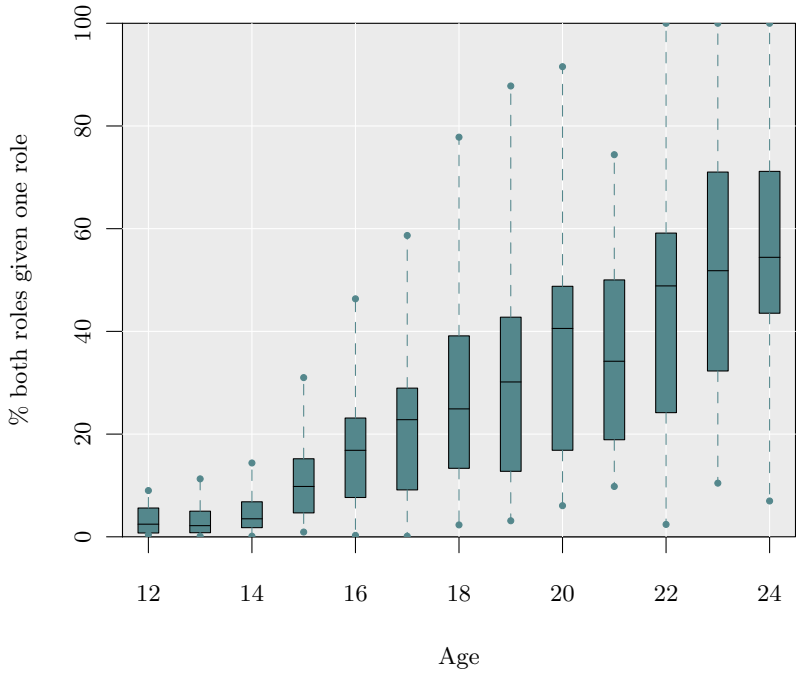
3.2 Figure 3, Male In Union by Enrollment, single ages

For males, again we are unfortuneately limited to union status. Figure 3.2 does not contain many surprises, except for too many zeros in the upper ages; we need to look at tables to see which countries are the culprits, and possibly evaluate the data. Differences in union status based on school attendance are clear in all ages, but only strongly separate starting around age 17. Point made.

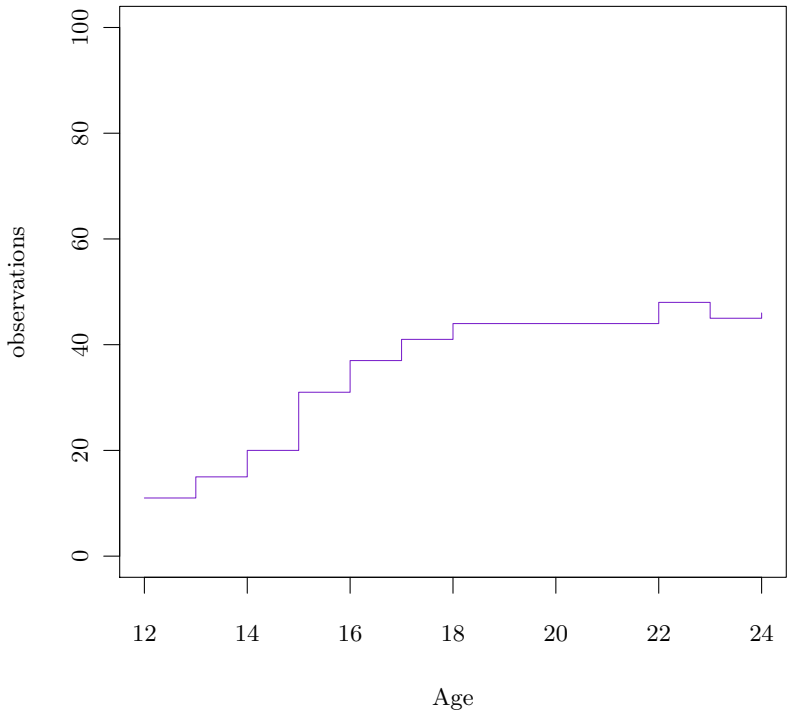


3.2.1 Figure 3b, Male In Union by Enrollment, single ages, all observations

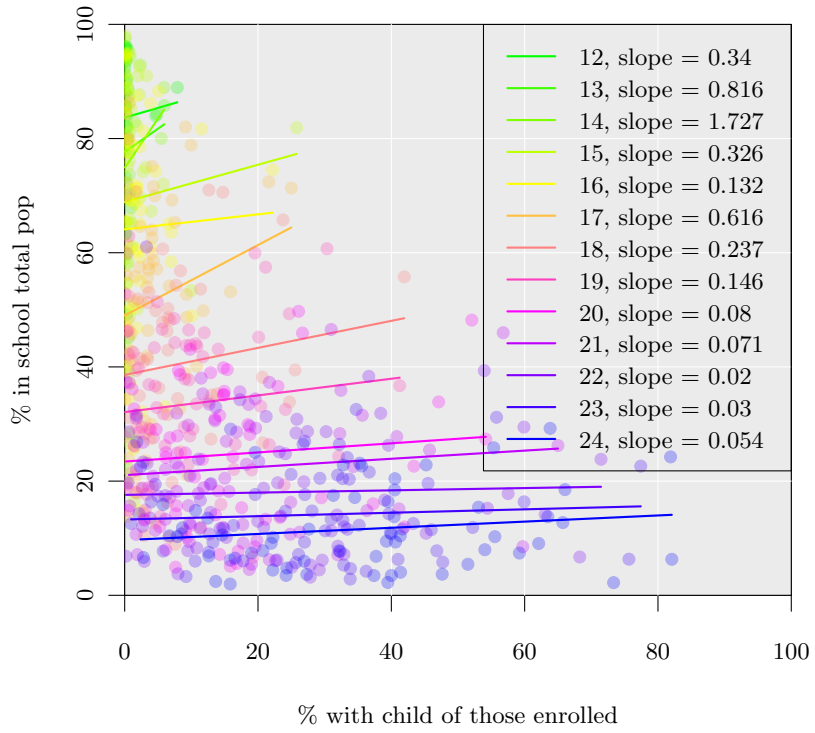
Figure 3.2.1 displays the observation distribution behind the prior boxplot. Strange things, probably sample size issues, are happening in the lower ages. Countries with greater than 10% of males in union and in school are Senegal (age 12 and 13), Rwanda (13); and out of school age 12 but higher than 10% married are Malawi, Rwanda and Senegal, and at age 13 the same countries plus Thailand. The consistency between ages almost suggests that this might not be a mistake, but I'm guess that these proportions are based on much less than 50 individuals (12-year old males in and out of school).



4 Figure 4 boxplot, Females, simultaneity parenthood and union, given either parent or in union

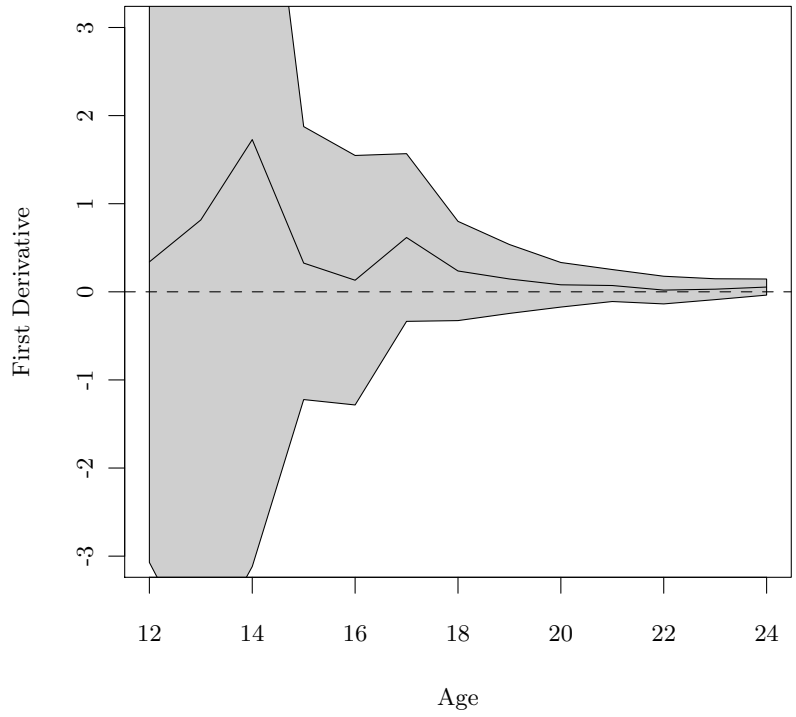


4.1 Figure 4b, observation counts for Figure 4

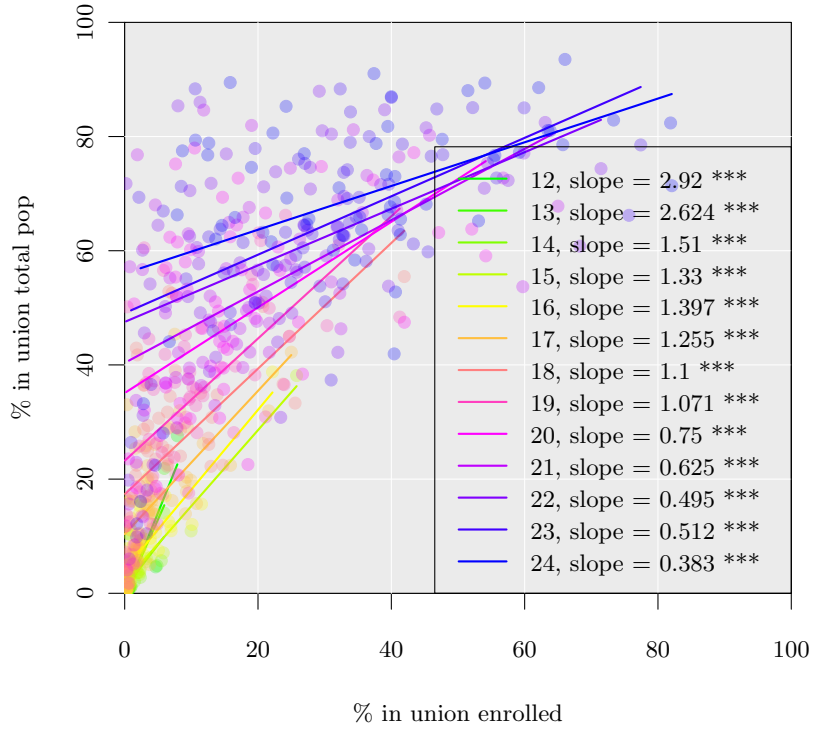


5 Figure 5 scatterplot, Females enrolled total population vs with children among attending

No time yet to comment on Figure 5.

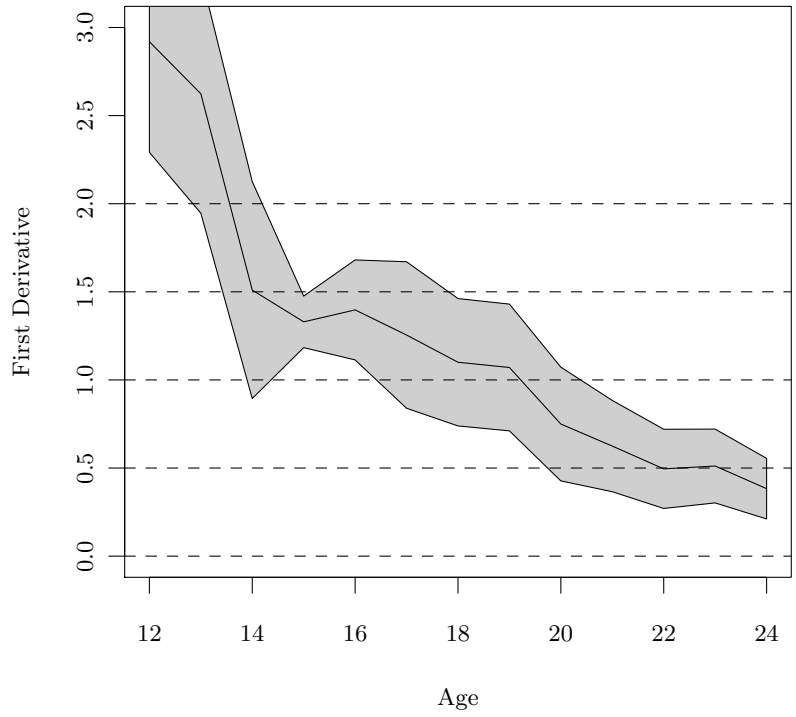


5.1 Figure 5b, change in slope of age-specific OLS

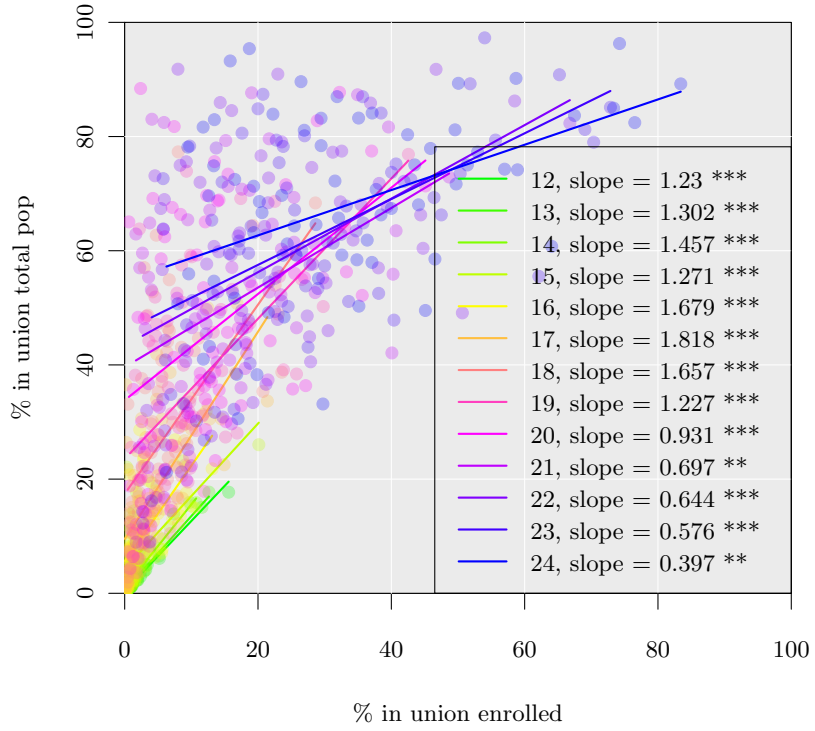


6 Figure 6 scatterplot, Females with child and in school vs with child total population

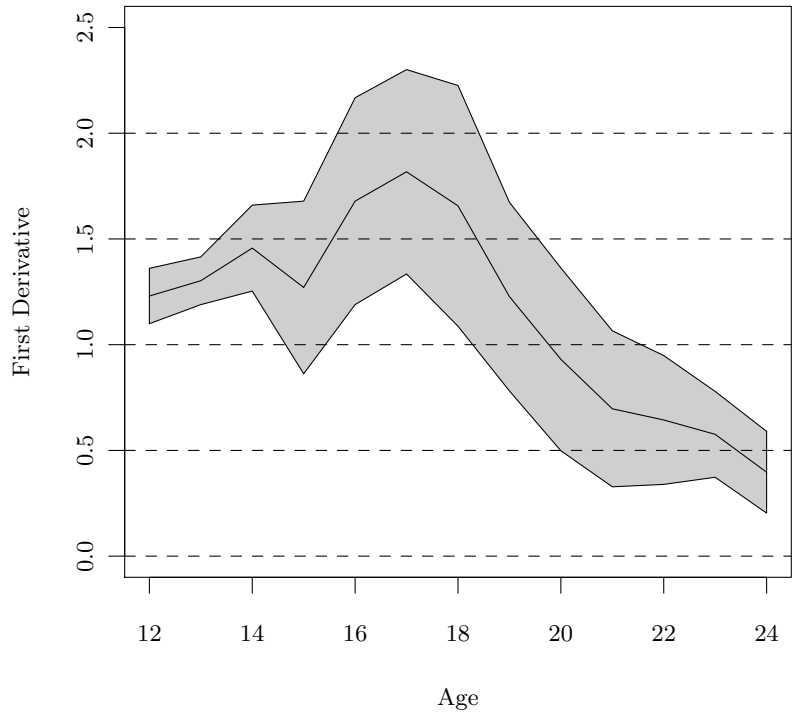
In the prior meeting, Albert mused that proportions in union while in school probably have a close relationship with proportions married in the population as a whole. Let's take a quick look. Figure 6 shows a strong and significant positive relationship between the overall porportion in union and the proportion in union and in school. Next we'll see how the relationship changes over age, and how linear the relationship really is.



6.1 Figure 6b, change in slope of age-specific OLS



7 Figure 7 scatterplot, Females in union and in school vs in union total population



7.1 Figure 7b, change in slope of age-specific OLS