Three major and one minor issue need to be worked on for the women’s labour force participation paper.

## Dealing with the omitted variable bias

The fundamental assumption in any regression analysis is that the error terms are uncorrelated with the explanatory variables. The error terms are everything that is not explained by the explanatory variables. When there is some covariation between an explanatory variable and the dependent variable that is due to the same driver, and this driver is not captured in the explanatory variables (it is omitted), then there is a bias to the estimated coefficients of the model. This bias can become very large. Such a variable with a common driver is called endogenous variable.

We develop a nice theoretical framework that explains that the perceptions women have about the labour market are to some degree influenced by the same drivers as the labour market outcomes themselves. At the same time, there is no way we can claim that we capture all the influential factors. Hence, we are subject to omitted variable bias.

The usual procedure to deal with such a bias is instrumental variables, for example used in two-stage least squares. The idea is the following: you regress the endogenous variable on instruments that can predict this endogenous variable, but which is uncorrelated with the dependent variable. Then you predict the endogenous variable, and use the prediction in a second stage to run your regression of interest. This can also be extended to multiple endogenous variables.

The key is to find an instrument that is meaningful, but uncorrelated with the dependent variable (labour market participation). Could you write a discussion on this problematic, going through the variables and identifying whether we could instrument things like the level of education and the preference to have a paid job?

In case we cannot find a useful instrument, we have to leave these variables out of the regressions, which would really be a pity.

## Using non-parametric methods to estimate the model

We always struggled to find a good functional form for the model: which country group to run, which interactions to set up. As it turns out, our problem is an excellent candidate to run a non-parametric estimation. In such an approach, we do not need to specify a functional form. Essentially, non-parametric regression identifies a density function for the data conditional on all the explanatory variables that makes it most likely that the data were actually generated by such a density function. This approach is very flexible, and also removes our problem of small cell size (for instance very little atheists in Arab countries.

This approach will not produce any coefficients. However, even with the parametric approach we used before the coefficients were useless, we had to compute marginal effects to tell a story. This is the same here, so no change.

I will work on setting up the estimation of the non-parametric model. This has to be done in R, as all the routines are programmed for R. This is free software, so in case you want to check it out, you can.

Non-parametric estimation removes the problem of finding a functional form, but the omitted variable bias remains, so we still need to figure out how to deal with that problem, using instrumental variables if possible.

## Identifying the story that we want to contrast

We need to identify the story we want to investigate. What are our variables of interest? We can look at the differential impact of certain variables across certain groups. Note that with non-parametric estimation we can show the whole distribution, meaning we can see how say access to transportation changes for with every year of age.

We had the idea of looking at the differential effect of poverty over age, but we can do more things. The only thing we might not be able to do is the effect of preferences, unless we find an instrument for it.

## Setting up a version control system so we can effectively work together on the project

Finally, it would be best to set up a version control system. This works very easy. I will set up an account on GITHub where the whole project is saved. The data in this account is synchronized with a folder on your computer. After working on files, you can commit them to the server. When I worked on files, you can synchronize and see the changes. In fact, all changes done over time can be tracked like this, so you could also go back to a previous version, without having many file names.

This works best with a software that has an integrated GITHub environment. TexStudio has a built-in feature to check in and check out from a version control system, where you can also look at differences between the online and local version.