Comments and tasks - Notes from Wendy July 2025, REVISED v2 17 July 25.

1 We need some EDA and to have a chunk of code that selects the age group 15 to 44 inclusive, in the analysis step which follows it.

Anthony can implement the age limits at end of step 1. I recommend age 15 to 44. %>% (age>14 & age<45)

* Keep these variables and count sub-groups if necessary:
* DONE; CHECK THE VAR NAME: Age Sex Rel. Let Sex female be 1 and sex male be 0. Let rel be HHHead = 1 and other be 0, rename this variable as Ishhead.
* Highest Education – Labourforce- Unemployed –need to carefully omit those who are not in the labour force..
* Create a vector Personnum which has the same values as Line. Keep line.
* Nested Calculation:
  + If sex=female and Ishhead=1 then make a new vector for HasFemHd.
  + Then spread this coding of HasFemHd to all members of that household . You may use Personnum for this.

In workshops, ambitious participants can use a different age scope for their exercises.

They may add the variable HasFemHd to their regression to improve fit and re-test the spatial component.

It is now agreed that we will assume all the adults in this dataset are in the Labour Force [LF]. I don’t think that’s reasonable. I think you mean people eligible to work, from which there are normally 3 subcategories:

Unemployed as you know;

Inactive which might be ill or retired, student, and many women; and

And Active which means Labour Force (LF).

We’re going to redefine LF to include the unemployed people in the LF. !! TYPICAL U.N. definitions are applied this way.

Remove the Inactive from the dataset during Step 2. This will make unemployment a binary:

Unemployed or Activew. But, this is subject to two separate Data Generation Processes. (DGP)> Gelman says that we should model each DGP not just an outcome. So for the

* Unemployed DGP we will expect [as independent variables] the Spatial units to dominate as an explanatory factor – they parallel policies and norms/institutions of unemployment. As a substitute one often uses Rural/Urban, avoid this! Because, Rural/Urban overlaps with our Spatial Units which are now potentially explanatory on Unemployment.

And

For the Labour-Force Active, but not Unemployed: They must be old enough to work; either sex can work but norms differ, [working toward a list of Independent Variables]; we put Female as Sex to predict a – sign; and age-group in five-year bins acts as a proxy for life stages. So for example the predictions work as follows:

If >20 years, will marry soon, will want to work (IF MALE) and will want to NOT WORK [out of data set] if FEMALE;

If >25 years, has borne kids typically, will want to work (IF MALE) and will want to NOT WORK (EXITS THE DATASET) IF FEMALE. SO we adjust for sex bias in each age group (INTERACTIONS).

If >44 years, kids left home, all ambiguities so we can leave these out of dataset.

3. Regression format can be: Put in these variables Age-group in 5-year bins 15-19, 20-24, 25-29, 30-34, 35-39, and 40-44. Interact sex by each Age group to maximally adapt for the exiting of women due to childrearing.

Next, for those women in the dataset, the risk of working is higher as they have not exited, so we allow MPCE (divide household annual spending by household raw size) to enter the regression in 3 stages: LOWEST< MIDDLE HIGH, with HIGH being the Reference case.

We again interact MPCE dummies with sex.

4. Now move to the geographic predictions, error term and BYM2.