**Notes on reconciling Python and R model results**

8/5/2020 Chris Zhang

|  |  |  |
| --- | --- | --- |
| Priority | Assigned to | Description of issue |
|  | Luke | low\_wage in R ACS are decimal (wage?). It should be binary. See data dictionary for post-sim ACS. |
|  | Luke | R ACS has variable names like ‘married.1’ likely duplicated column from merging in R code. |
|  | Luke | First column of R ACS has no column label – please verify what this column is or if it’s needed at all. |
|  | Luke | R ACS has empsize simulated as exact numbers. Need to convert to code 1-6 as CPS. See data dictionary for empsize definition. |
|  | Chris | [fixed 08/05/2020] ~~Python model cannot replace results using same seed. Replication is fine with R model in GUI.~~ |
|  | Chris | [fixed 08/05/2020 – updated cps\_clean\_[year].csv, 3 files] ~~In Python, empsize predicted to be 0 (CPS code NOT IN UNIVERSE) and will be excluded from eligible workers (empsize>=1). Correct this in CPS/ACS imputation.~~ |
| urgent | Luke | R does not store intermediate cleaned FMLA/ACS for sim. Need fmla\_clean\_[year].csv in data/fmla/fmla\_2012, and ACS\_cleaned\_forsimulation\_2016\_[st].csv in data/acs. These are respectively the cleaned input FMLA and ACS data files for model training and predictions. |
|  | Chris | take/need\_types in ACS are different between Py/R. NT check   * 1. cleaned FMLA input data   2. cleaned ACS input data   3. NA handling   4. classifier (logit GLM) |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |