Here we try to explain how Cengiz et al. (2019) create their main employment dataset. This is based on the do file state\_panels\_cents\_new\_ QJE.do, which is the first code in the series of codes that they use to create their employment dataset. Our main focus is on how they build the outcome variable the regression 1 of their paper E\_{sjt}/N\_{st}; where E\_{sjt} is “the employment in $0.25 wage bins j in state s and at quarter t”, and N\_{st} is “the size of the population in state s and quarter t”.

To do so, we introduce three variables that understanding them helps us to understand how the outcome variable is being built:

1. totalpopulation: Sum of earnwt for each quarter-state before dropping any observation (N\_{st})
2. totpopcount: Sum of earnwt for each quarter-state-bin after dropping self-employed, imputed, and nonpositive wages (E\_{sjt})
3. countall: Sum over j of the totpopcount (sum of the E\_{sjt} over the j)

earnwt is the weight of each observation in the survey. This is the description of the earnwt in the CPS Labor Extract 1979 - 2006 (Feenberg and Roth 2007):

"Earnings weight for all races. Used for tabulating earnings-related items. Since the CD-ROM includes all persons asked earning questions, this sums to the total population each month and 12 times the population for each MORG file."

Here are some rows of the data, where we observe some values of these variables for California in the last quarter in the data:

A screenshot of a computer

Description automatically generated

First, we observe that totalpoualtion is around 31 million which is the civilian noninstitutional population (population of age 16 years and above) of California. Second, if nothing has been omitted before creating the E\_{sjt}, we would expect totalpopulation and countall to be the same. However, this is not the case and totalpopulation and countall are different. The difference is explained by dropping:

1. Self-employed persons
2. Imputed wages
3. Nonpositive wages

If one comments out these lines in the code, there will be no difference between totalpopulation and countall:

* Dropping self-employed:

*drop if (class==5 | class==6) & year<=1993*

*drop if (class94==6 | class94==7) & year>=1994*

*drop class class94*

* Dropping imputed and nonpositive wages

*replace wage = . if imputed == 1  
gen logwage = log(wage)  
drop if logwage == .*

References

Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zipperer. 2019. “The Effect of Minimum Wages on Low-Wage Jobs\*.” *The Quarterly Journal of Economics* 134(3):1405–54. doi: 10.1093/qje/qjz014.

Feenberg, Daniel, and Jean Roth. 2007. “CPS Labor Extracts 1979 - 2006.”