This document is contains notes from my process of using the Eeckhout et al code to replicate Figure 5 from their paper. The purpose of these notes is to figure out how we can do an analysis for 1970, 1980, 1990 and 2000 which will closesly mimic their procedure.

Specifically, we are going to be looking at the following samples:

* 2000 5% sample
* 1990 1% metro sample
* 1980 1% metro sample
* 1970 1% metro sample

To determine any changes that may need to be made in adapting the code for different years (or sensitivity checks to be performed), I will look at the IPUMS documentation for each variable we are using to see if there are difference between the codes for different years, and make sure to read the ‘comparability’ tab thoroughly.

Notes from data.do:

Sections: CBSA labels, PUMA to CBSA Correspondance, and CBSA population.

* Since CBSAs are not available going back in time, we will be using metro areas rather than CBSAs. A drawback to this is that we are essentially looking at pooled cross sections instead of a panel, because the MSA identifiers are not consistent across time like the CBSA identifiers are.
* They assign PUMAs (and hence houses/workers) to CBSAs by finding the CBSA that the largest portion (by fraction of households) of the PUMA lies in. PUMAs that do not belong at least 1/3 to any particular CBSA are dropped. This will exclude some of the population towards the perimeter of a given CBSA, as well as not accounting for uncertainty in the case of a PUMA which has significant portions lying in 2 different CBSAs.
  + Whereas we will assign observations to metros by using the IPUMS metro area codes, which also exclude some of the population on the periphery, as is well documented on the IPUMS website.
* For any observation that is not identified to a CBSA, the value of cbsa==. So in processing the data for other samples, I need to make sure that I similarly have metarea==. for observations not identified. This may be as simple as replacing 00 with missing, might be more complicated.
* They get CBSA populations from a list compiled by someone at the Census. We will get populations by summing the `perwt` variable within each `metarea`. I’ve compared the values I get using this technique to some published figures, seems like a reasonable way to do things.

Section: Hedonic Price Indices

* They drop for Group Quarters codes 3, 4 and 5. The only difference will be that in 1970 we will be dropping households with 5-9 individuals unrelated to the head of household, while we will keep them in all other samples. Could test sensitivity by dropping GQ==2,3,4, and 5.
* They drop for Farm ~=1. There are only 2 codes for farm in all of the samples we will be looking at. In 1970 a farm was either 1) a household on 10+ acres that yielded $50+ in produce, or 2) a household on fewer than 10 acres that yielded $250+ in produce. In other samples, a farm was any household on 1+ acres that yielded $1000+ in produce in the prior year. This is a minor difference and there is nothing we can really do to check sensitivity. Our assignment procedures for 1980 onward already ‘skew toward core populations and omit outlying communities’ to quote IPUMS, so farms are probably not much of an issue.
* For unitsstr, they drop codes 01 and 02. This includes boats, tents, mobile home or trailer, other. Code 02 is not present in the 1970 sample, but otherwise things are very comparable. One thing hardly worth noting: “In 1980 and 1990, the instructions said to count a mobile home or trailer as a detached house if a permanent room (excluding porches and sheds) had been added. In the 1970 and 2000 censuses, the ACS, and the PRCS all mobile homes were grouped together in a separate category.”
* To transform the person-level dataset to a household-level dataset, they have a whole section of code, whereas I was just doing keep if pernum==1. The advantage of their method is that they construct a variable recording how many members of the household were in the dataset, which might be useful. I altered the code to keep first values out to `incearn`. May need to be further modified if more variable are added or variable order is switched for some reason.
* When they drop observations (households) not in a known CBSA, this drops 110,306, leaving only 991,697. Is this comparable to how many households I lose when I drop metarea==0?
* For the rent variable, they replace rent==0 with missing values. Also, they are using the plain contract rent variable, when I think it would be more reasonable to use the gross rent variable, based on the IPUMS documentation.
* The number of rooms is top-coded in earlier samples, so replace values greater than 9 with 9. I’m actually not sure why they bothered to include this since they are only using the 2009 data, but I’m glad they caught it.
* They define variables to delimit the samples to be used in the regressions. They base this on the criteria of: being in an identified CBSA, paying cash rent, and having a positive value of rent. One sensitivity test might be to change from looking at `ownershpd==22` to `ownershp=2`, what’s the deal with the restriction to “cash” rent? Oh nevermind, looking at the documentation and questionaire text it seems like a correct restriction.
* For the regression itself: left hand side variable of log rent should be fine across all samples. The right hand side variables are as follows:
  + `rooms`. As mentioned previously, this is top-coded at 9 for most samples, so we have a step to deal with that.
  + `builtyr2`. This deals with the age of the structure, but is only available beginning in 2000. `builtyr` is avialable for 1970 through 2000. Since we don’t care about the coefficient on this variable, and just want to condition on it, I see no issue with using `builyr2` for 2009 and 2000, and `builtyr` for earlier samples.
  + `unitsstr`. The only inconsistencies here are noted above. Since we are not trying to do any panel-data price indices where we are sharing information across years, I don’t see any issue.
  + NOTE: they use the `bi.` notation to specify base categories. Might need to change these or replace with `i.` for other decades.
* REPLICATION SUCCESS: all coefficients are coming out exactly the same in the CBSA regression as what is published in table B2.
* I have deleted parts of their code that estimated fixed effects for PUMAs instead of CBSAs, since this will not be possible for earlier samples. It might be good to change all of the references to CBSAs to METAREAs when I adapt the code for eaerlier samples.

Section: Wage and Skill from ACS data

* They use (and our analysis will use) the `incwage` variable for an individual’s income. This variable is top-coded, with different top codes in each sample, so handling that may require some changes to the code for each decade. Rounding is also slightly different between years, don’t think this an issue. Overall, comparability tab from IPUMS suggests there is nothing to worry about.
* Might want to try using `incearn` instead of `incwage`, as this includes farm and (more importantly) business income. This measures something different for sure, as you will probably get more people at the top end of the distribution this way, but worth checking out. `incearn` begins in 1990, but it is just the sum of incwage with incbus and incfarm, which can be done going back to 1950.
* They only look at full time, full year workers. This involves filtering on 3 variables:
  + Labforce==2. This is available and consistent beginning in 1940 (but labor force participation was definied/asked differently 1930 and prior). To be clear, this question filters on “LAST WEEK, did this person work for pay or profit”. Might be better to use the variable `workedyr`, which asks if they worked in the previous year at all.
  + Wkswork2==5|6. This is available going back to 1940, and filtering to values 5 and 5 gives us people who worked 48 to 52 weeks last year. Measures weeks that people did any week, or paid absence, not “full time equivalent week”.
  + 36<uhrswork<60. This filters to ‘full-time workers’ based on their usual number of hours worked per week last year. `uhrswork` is not available for 1970 and prior, so we will have to use `hrswork2`, hours worked last week. If you think that hours worked last week is a noisy but unbiased measure of usual hours worked, then that shouldn’t really pose an issue. It’s hard for me to think of any reason why their would be a problematic systemic bias with hours worked last week, would probably need to look at the sampling design to see when different kinds of people were sampled. From IPUMS Comparability tab: “because the census was taken at different times of year, the reference week ("last week") varied somewhat from census to census and within censuses. In 1940, the reference week was March 24-30. In later years, the reference week was simply "last week," but since the census was administered over a period of time, "last week" was not the same week for every respondent. The holidays of Passover and Easter were part of some respondents' reference weeks in U.S. and Puerto Rican census samples in 1980 and 1990, which probably meant that some respondents reported fewer total work hours than they would have for most other weeks. Census Bureau studies indicate that the number of people who worked only a few hours may be understated, since many such people do not consider themselves to be working.” According to some 1980 Census documentation, April 1 is Census Day because that’s when there is the best chance of people being at home and being able to reach them.
* To calculate their wage variables, they divide by different numbers for people who worked 48-49 vs 50-52 weeks last year. This gives WEEKLY WAGE. Not obvious to me why you wouldn’t also want to look at HOURLY WAGE by dividing by usual hours worked.
* They drop the lowest 0.1% of wages to deal with outliers, should probably carry this forward in our work.
* They keep track of birthplace, for reasons not yet clear to me.
* When merging the PUMAs to CBSAs, they lose observations in one part of New Orleans, and some other rural (as in metaread==0) PUMAs. However, there are still notable differences between the `cbsa` variable we end up with compared to the `metaread` assignment.
* They drop for `metro<=1`, which is unidentified or not in metro area observations. `metro` availability and code meaning changes a lot over time. Ultimately, since we are going to use metarea as reported in the IPUMS sample, this should not be an issue for us.
* They drop if missing(`cbsa`). Similarly, we will drop if missing(metarea).
* They drop if the `wage` variable they constructed is missing. We should keep this line.
* I’ve deleted the portions of their code that used the PUMA housing price indices, otherwise the skill section of their code seems appropriate.

Section: Figure 5

* They start by bringinig in the skill data file from the last step.
* They drop observations in the CBSAs with population less than 100,000 (which due to the assignment strategy of PUMAs to CBSA is very low amount of observations).
* They then define a group variable based on city population. We will want to alter this code to use percentiles of city size distribution instead of the fixed cutoffs of 1M and 2.5M.
* Their code is using aweights for the quantile regressions, this is no longer allowed in newer versions of STATA.

REPLICATION SUCCESS: some quick and dirty cutting, pasting and editing of their section6 code into my old 2009 code produces a graph that looks VERY similar to their results, using METAREA classification instead of CBSA classification.