**RAW MATERIALS TO HELP YOU WRITE THE METHODS SECTION**

1. **What Data Will We Use?**

* 1920 US Census Data linked to 1940 US Census Data linked to mortality records form the Social Security Administration’s Death Master File (SSDMF) and its NUMIDENT file.
* From the 1920 Census, we get demographic info (e.g., age, race, sex, place of birth), socioeconomic info about the family (e.g., parents’ occupation), and geographic location.
* From the 1940 Census we get education, occupation, and income.
* 1920 and 1940 Census data are “complete count” data (meaning it’s not sample data … everybody in the country is in the file) and comes from IPUMS.org (which you should go check out).
* SSDMF is the Social Security Administration’s file for keeping track of who has died. You can read more about it here: <http://www.sciencedirect.com/science/article/pii/S1047279716304884>
* The NUMIDENT is the Social Security Administration’s “claims” file … when people apply for benefits. It has — important for our purposes — people’s parents’ names
* You should note that the SSDMF really only covers deaths from about 1970 through 2013. The public version of the NUMIDENT file covers about the same range of years but has way fewer deaths. People can appear in one or the other or both. We combine them to maximize the number of deaths we observe.
* We start with people age 0 to 10 in 1920. There were about 25 million people age 0 to 10 in 1920.
* Of course, people have to have survived to at least 1940 (age 20 to 30) to be in our final data set. And, we don’t observe deaths between 1940 and 1970. So, we miss people who died in that middle period (1940 to 1970, or ages 20-30 to 50-60).

1. **How Are the Records Linked?**

* We attempted to link (almost) every 0-10 year old boy in the 1920 Census to his record in the 1940 Census and to his record in Social Security Administration mortality records
* We did NOT try to link people born outside the US. Linkage rates are just too low.
* We follow the probabilistic, machine-learning based approach spelled out by Feigenbaum (2016) … see https://hdl.handle.net/2144/27526
* Basically, we link on names. Specifically, we use Jaro-Winkler distances to quantify how similar two names are. Some references to cite:

Jaro, M. A. (1989). "Advances in record linkage methodology as applied to the 1985 census of Tampa Florida". Journal of the American Statistical Association. 84 (406): 414–20. doi:10.1080/01621459.1989.10478785.

Jaro, M. A. (1995). "Probabilistic linkage of large public health data file". Statistics in Medicine. 14 (5–7): 491–8. doi:10.1002/sim.4780140510. PMID 7792443.

Winkler, W. E. (1990). "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage" (PDF). Proceedings of the Section on Survey Research Methods. American Statistical Association: 354–359.

Winkler, W. E. (2006). "Overview of Record Linkage and Current Research Directions" (PDF). Research Report Series, RRS.

* However, we don’t just compare EVERY boy age 0-10 in the 1930 Census to EVERY man age 20-30 in the 1940 Census (or to every boy born between 1910 and 1920 who is in mortality records). Instead, we “block” … we only compare boys in 1920 to boys in 1940 who shared two (theoretically) immutable attributes: The state they were born in and the year (within plus or minus 3) they were born. That is, we don’t compare “Richard Nixon” (born 1913 in California) in the 1920 Census to every boy in the 1940 Census; we only compare him to boys born in California between 1910 and 1916. This vastly reduces the computational burden.
* We then compile, for every boy, a list of “plausible” matches. There are typically a dozen or more “plausible” matches.
* The “plausible” matches are those with a Jaro-Winkler score of at least 0.8 (and, of course, who match on state of birth and year of birth plus/minus three). So, Richard Nixon might have plausible matches like “Richard Dixon” and “Richrd Nixon” and “Richert Nixon” and “Richard Nixon”
* How do we then know which “plausible match” (if any) is a valid link?
* Here, we use training data to teach the computer how to recognize which, if any, of the “plausible” links ae correct.
* First, we link a random subset of 2,500 cases by hand. We use everything we know about the boys from the Census and mortality files, and we use the powerful databases and search features at Ancestry.com. (Just like you did in your assignment!)
* Then, we use a probit model (which is just a fancy regression model) to predict which of the “plausible” matches is the right match (where “right” is determined by hand linking). The model predicts which “Richard Nixon” is a match based on the Jaro-Winkler score, age, geography, race, and a few other things.
* We only estimate this prediction model on HALF of the 2,500 cases in the training data.
* We then evaluate how good the model works on the OTHER half of the training data; how often does the model make a correct link? How often does it make an incorrect link? (Again, we know if it’s getting the “right” answer by comparing to the hand-linking results)
* Based on this evaluation of the model, we re-calibrate it … fine tune it so it does a better job of predicting the right answer. Our goal is to minimize the rate of false positive links (i.e., we want to prevent linking somebody to the wrong person at all costs) while maximizing the total number of links (so we have a large sample size). That is, we want to link as many people as we can while being very conservative so as not to link to incorrect people.
* Once we have optimized the machine linking algorithm (i.e., the probit model), we apply it to the entire huge dataset to get predicted matches for all boys. We can be confident that the number of false links is low, and the total number of links is as high as possible.
* Note that we declare a link if any only if (1) there is at least one possible match that has a high probability of being correct and (2) the “goodness” of the most likely link is much better than the goodness of the second best link.
* Keep in mind: We really do all of the machine linking TWICE. First, we link 1920 to 1940. Then, we link 1920-1940 to mortality records. Thus, everything above (i.e., training data, models, etc) gets done twice.

1. **How is The Sample Selected / Restricted?**
2. Number of 0-10 year old US-born people in the 1920 Census: 25,241,189 boys
3. Number who are men: 12,765,719
4. …and who are kids or step-kids of the household head 12,498,173
5. …and who are fully linked to the 1940 Census & death records 1,776,098
6. …and who have valid birth & death dates 1,776,076
7. …and who have 2+ linked boys in the household 359,732

* “a” only includes US-born people because foreign-born people are so hard to link
* “b” limits the data to men … necessary because women are hard to link because they change their names (usually) between 1920 and 1940.
* “c” restricts the sample to children of the household head. Why do this? We don’t want to assume that two cousins are brothers, or that two boys of different fathers living in the same house are fathers, etc. We just want people who are brothers.
* “d” restricts the sample to people for whom we see 1940 info (i.e., education, occupation, income) and for whom we can get mortality data
* “e” throws out the few cases that have missing info on death date from mortality records; not sure why those are missing
* “f” is necessary to get two — and only two — boys in households. We throw out boys who do not live with any brothers age 0 to 10. When there are more than two brothers in a household who are age 0 to 10, we randomly select two of them.
* So in the end, we have pairs of brothers with complete information.
* Our final sample includes 359,732 boys in 179,866 households. That is, we have 179,866 pairs of brothers with complete data.

1. **How Will We Do the Analyses?**

FOR QUESTION #1 (which is about the effects of having a particular kind of name on outcomes):

* To estimate the effects of the characteristics of names (e.g., how often people with that name are black or a northerner or male) on educational attainment, occupation, income, and longevity, we use a standard fixed effects specification for within-pair estimation:

*Mij*=  + *Nij*+ *Cj* + *ij*, (1)

* In Eq. (1), *Mij* is a continuous outcome measure (i.e., education, occupation, income, or longevity) for individual *i* (*i* = 1, 2) in sibling pair *j* (*j* = 1, 2, …, *N*); *Nij* measures attributes of the name of individual *i* in sibling pair *j*; *Cj* is a measure of unobserved contextual characteristics shared by both members of the pair (e.g., race/ethnicity, family socioeconomic resources, neighborhood attributes); and ε*ij* is a random individual-level error term that is assumed to be uncorrelated with the other explanatory variables in the model.
* In a within-sibling pair model, the unobserved shared components *Cj* are “controlled away” by modeling differences within pairs of brothers:

*M1j* – *M2j* = ∆*Mj*= ∆*Nj* + ∆*j*, (2)

* In Eq. (2), the ∆’s represent differences between variables for the *j*th twin pair (i.e., *N*1*j* – *N*2*j*).
* This approach eliminates the confounding effects of unobserved contextual characteristics (*Cj*)—since the vast majority of siblings experience the same racial/ethnic, family, school, and neighborhood environments while growing up.
* The model can be modified to include controls for individual-level measures (e.g., age) that are not shared by siblings but that may influence outcomes.

FOR QUESTION #2 (which for most groups is about how the effect in Question #1 depends on race or whether you are from the north/south or whether you are an immigrant):

* To understand whether the effects identified in Eq. (2) above differ for white and black males (or, for northern and southern males or whatever) we will estimate separate models for sibling pairs who are (1) white and then (2) black.

1. **What Problems Do We Anticipate? How Will We Overcome Them?**

* First, we know that we are not able to link everyone. This raises a possible concern: Maybe the people we could link are not a representative subset of the population? And thus, maybe the results we get don’t generalize to the whole population (i.e., are wrong)?
* To overcome this concern, we will take advantage of the fact that we have complete information about all boys—linked and unlinked—from the 1920 Census. Thus we can compare linked and unlinked boys with respect to race, state of residence, rural/urban residence, parental literacy, parental occupation, and family size.
* Second, we can never observe age at death for people who die between 1920 and about 1970 (i.e., between ages 0-10 and 50-60.
* However, we can (and will) show that relatively few people die before these ages and that our data capture most deaths after about 1970.