Finding Fake People in the United States Census

ALLEN OTTERSTROM

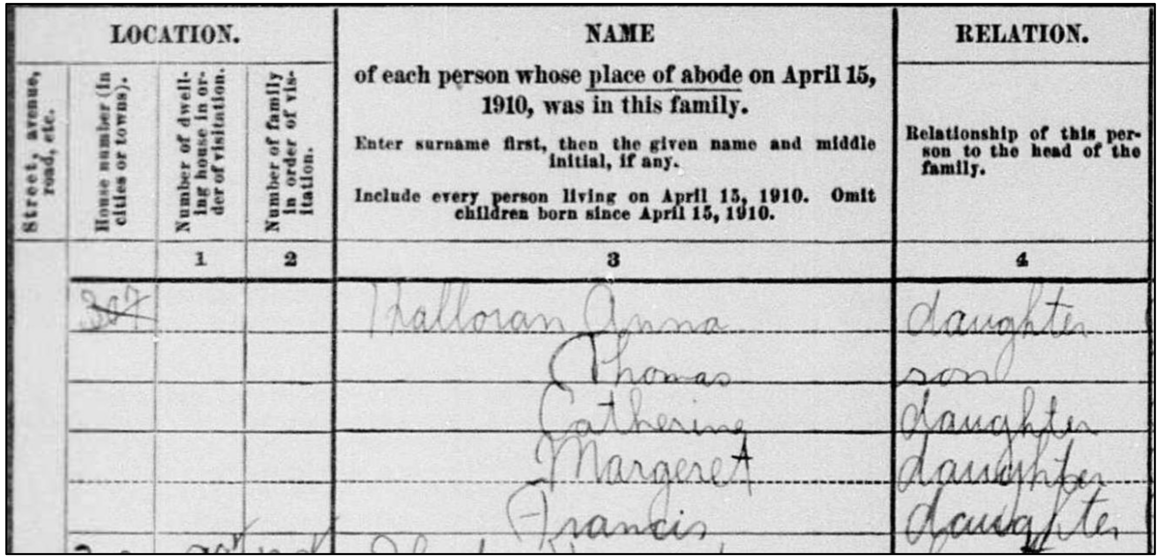
BYU RECORD LINKING LAB

Questions

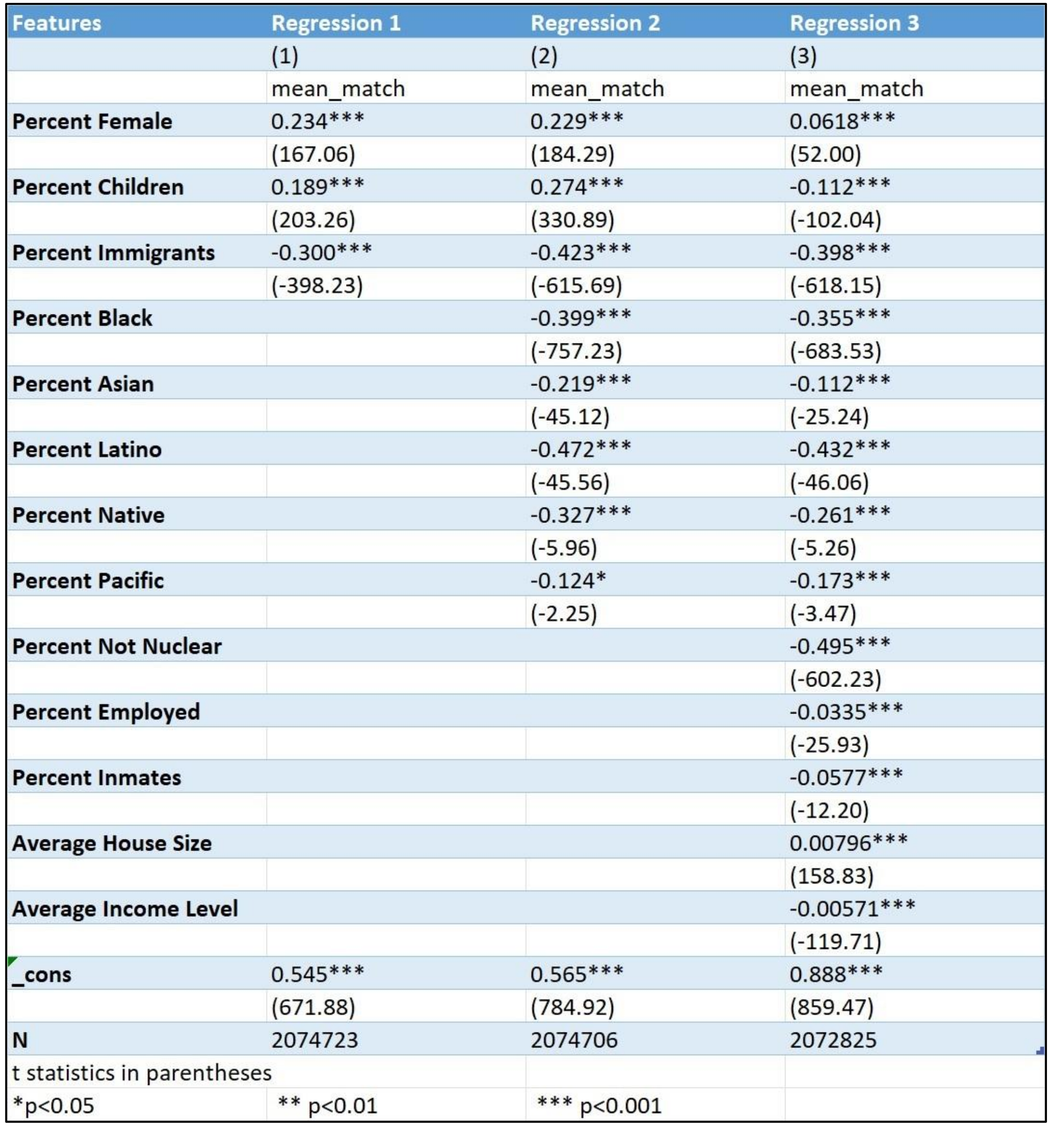
Are there “fake people” in the United States Census?

How do we find these “fake people”?

Why is this important?



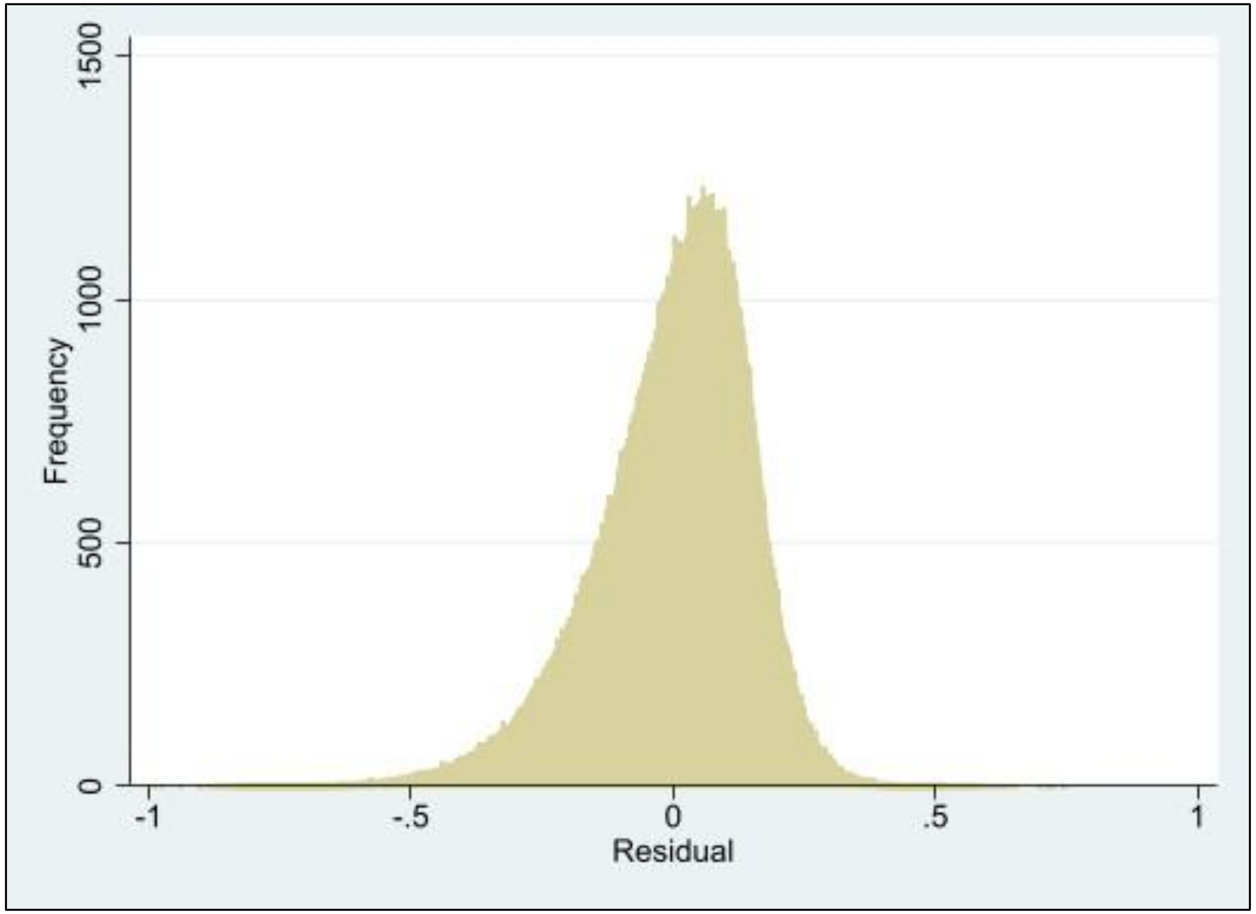
LOCATION.				NAME	RELATION.
Street, avenue, road, etc.	House number (in cities or towns).	Number of dwelling house in order of visitation.	Number of family in order of visitation.	of each person whose place of abode on April 15, 1910, was in this family. Enter surname first, then the given name and middle initial, if any. Include every person living on April 15, 1910. Omit children born since April 15, 1910.	Relationship of this person to the head of the family.
		1	2	3	4
	~~307~~			Halloran Anna	daughter
				Thomas	son
				Catherine	daughter
				Margeret	daughter
				Francis	daughter



Features	Regression 1	Regression 2	Regression 3
	(1)	(2)	(3)
	mean_match	mean_match	mean_match
Percent Female	0.234***	0.229***	0.0618***
	(167.06)	(184.29)	(52.00)
Percent Children	0.189***	0.274***	-0.112***
	(203.26)	(330.89)	(-102.04)
Percent Immigrants	-0.300***	-0.423***	-0.398***
	(-398.23)	(-615.69)	(-618.15)
Percent Black		-0.399***	-0.355***
		(-757.23)	(-683.53)
Percent Asian		-0.219***	-0.112***
		(-45.12)	(-25.24)
Percent Latino		-0.472***	-0.432***
		(-45.56)	(-46.06)
Percent Native		-0.327***	-0.261***
		(-5.96)	(-5.26)
Percent Pacific		-0.124*	-0.173***
		(-2.25)	(-3.47)
Percent Not Nuclear			-0.495***
			(-602.23)
Percent Employed			-0.0335***
			(-25.93)
Percent Inmates			-0.0577***
			(-12.20)
Average House Size			0.00796***
			(158.83)
Average Income Level			-0.00571***
			(-119.71)
_cons	0.545***	0.565***	0.888***
	(671.88)	(784.92)	(859.47)
N	2074723	2074706	2072825
t statistics in parentheses			
*p<0.05	** p<0.01	*** p<0.001	

Method:

1. Compare 1920 Census to 1910 Census to find the match rate per sheet
2. Use FamilySearch and Ancestry data to find demographic, ethnic and economic information
3. Run an OLS regression of match rate on that information



Residuals

- These residuals are found by making a prediction of the match rate using the results of the regression and then subtracting that prediction from the actual match rate
- Sheets with a residual closest to -1 are most likely where the potential fake people are