Inferring Hispanic Origin for Vital Records

Larry D. Sink, Population Division, U.S. Census Bureau

Population Association of America
Denver, Colorado
April 25-28, 2018

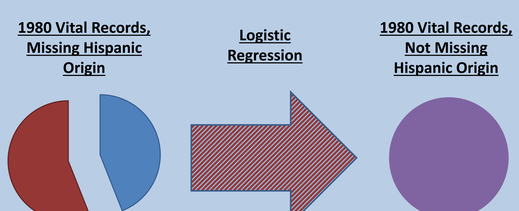
THE PROBLEM AND A PROPOSED SOLUTION

Problem

- Most states did not collect Hispanic origin information on vital records prior to 1990.
- Demographic Analysis (DA), which produces population estimates based on vital records, can thus only estimate the Hispanic population born since 1990.
- Time-series analysis of Hispanic vital rates is similarly limited.
- However, though only 22 states collected Hispanic origin information for most of the 80s, these states contained over 90% of the Hispanic population in 1990.

Proposed Solution

- Logistic regression can be used to predict Hispanic origin for those records from the 1980s which do not have that information, thus making another decade of Hispanic vital statistics information available for DA and other analysis.
- I use demographic information available on the vital records together with estimates of the population at risk to fit logistic regression models for Hispanic origin for those records where it is present.
- The results of these regressions are used to predict Hispanic origin for those records where it is missing.
- Birth records collect ethnicity information for the parents, but not the children, so I run models for the parents and use those results, together with census information, to predict the Hispanic origin of the children.



THE MODELS

Mortality Records

$$H = b_0 + b_1*AGE + b_2*HPCT + b_3*RACE + b_4*SEX$$

Where:
H = 1 if the decedent was Hispanic, 0 otherwise
AGE = age at death in single years, 0 to 85+
HPCT = percent of the population that is Hispanic in the decedent's 5-year age group for the decedent's county of residence
RACE = 1 if the decedent was White, 0 otherwise
SEX = 1 if the decedent was male, 0 otherwise

Natality Records

For Mothers:

$$H = b_0 + b_1*AGE + b_2*HPCT + b_3*RACE + b_4*FBRN$$

Where:
H = 1 if the mother is Hispanic, 0 otherwise
AGE = mother's age at birth in single years
HPCT = percent of the female population that is Hispanic in the mother's 5-year age group for the mother's county of residence
RACE = 1 if the mother is White, 0 otherwise
FBRN = 1 if the mother is foreign born, 0 otherwise

For Fathers:

$$H = b_0 + b_1*AGE + b_2*HPCT + b_3*RACE + b_4*PRBM$$

Where:
H = 1 if the father is Hispanic, 0 otherwise
AGE = father's age at birth in single years
HPCT = percent of the male population that is Hispanic in the father's 5-year age group for the mother's county of residence
RACE = 1 if the father is White, 0 otherwise
PRBM = probability that the mother is Hispanic

SUPPORTING DATA

County-level Hispanic Population

I created county-level estimates of the Hispanic population with age and sex detail to be used in deriving the percentage Hispanic variables for the logistic regression models. To do this, I used 1980 and 1990 Census counts of the Hispanic population at the county level along with the Census Bureau's national-level intercensal estimates of the Hispanic population in a process based on the cohort component method.

Using Origin of Parents to Estimate Origin of Children

Because the NCHS natality files contain Hispanic origin information for the parents but not the child, I create what we call a *kid-link file* to infer the origin of the children from that of the parents. This is accomplished through the following steps:

- Select household records from the 1980 Census containing Hispanic origin information on father, mother, and their children.
- Tabulate the proportion of cases where the children were identified as Hispanic from those records where the father was Hispanic and the mother non-Hispanic, and also from those records where the father was non-Hispanic and the mother Hispanic.
- Repeat this operation using records from the 1990 Census.
- For the intervening years, interpolate geometrically between the proportions from the 1980 Census and those from the 1990 Census.

PREPARING FOR THE 2020 CENSUS

Demographic Analysis will be used to evaluate the quality of the 2020 Census:

- A goal of the 2020 DA program is to expand the race and Hispanic origin detail of the estimates.
- This research shows how we could produce DA estimates for cohorts born before 1990.

RESULTS

Mortality Model

Parameter	Estimate	Chi-Square	Odds Ratio
Intercept	-5.024	140,000***	
AGE	-0.0167	40,710***	0.983
HISPCT	0.0694	304,200***	1.072
RACE	2.483	39,860***	11.98
SEX	0.1293	1,070***	1.138

***p<.0001

Using this model I obtained the probability that the decedent was Hispanic for every record (for those records for which origin was known the probability was one or zero). I summed these probabilities to the county level by year of occurrence, age, and sex to obtain estimates of the number of Hispanic deaths in that county.

Fertility Models

Mothers

Parameter	Estimate	Chi-Square	Odds Ratio
Intercept	-4.662	868,200***	
AGE	-0.0602	169,200***	0.942
HISPCT	0.053	1,568,000***	1.055
RACE	3.179	887,900***	24.18
FBRN	3.288	3,351,000***	26.87

***p<.0001

Fathers

Parameter	Estimate	Chi-Square	Odds Ratio
Intercept	-5.026	572,900***	
AGE	-0.0149	7,237***	0.986
HISPCT	0.0243	180,800***	1.025
RACE	1.868	188,200***	6.529
PRBM	5.032	4,800,000***	154

***p<.0001

Using these models I obtained the probabilities that the mother and father were Hispanic for every record. I used these results together with the kid-link file (described in the previous panel) to obtain the probability that the child was Hispanic, then summed these probabilities to the county level by year of occurrence and sex to obtain estimates of the number of Hispanic births in that county.

United States Census Bureau
U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov