STA440 Final Project

Matthew O'Donnell

2024-12-01

1. Introduction

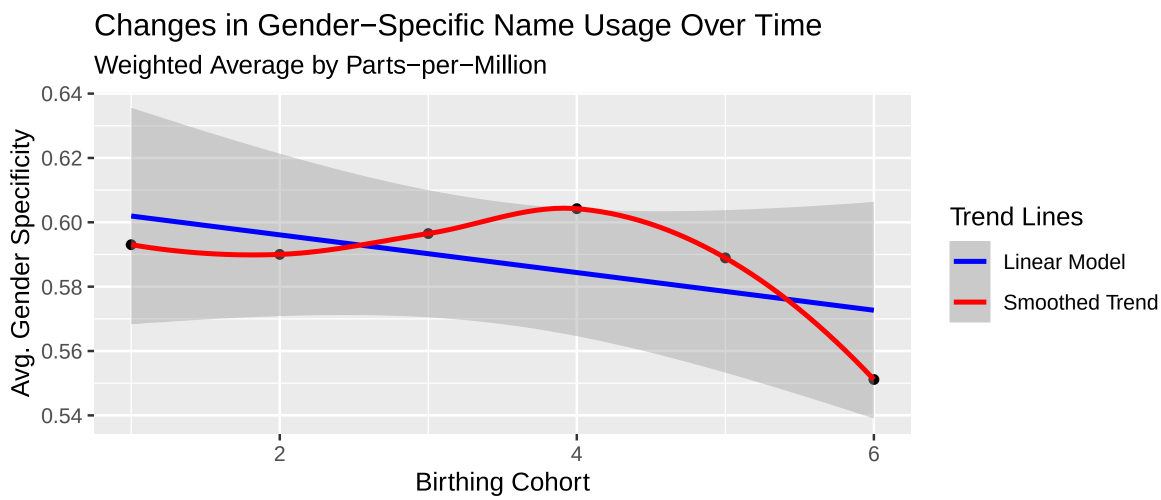
2. Methodology

2.1 Gender-Specific Name Usage Over Time

For each character in the dataset, a gender score is provided, which is computed as follows:

$$g_i = \frac{N_{\text{male}} - N_{\text{female}}}{N_{\text{male}} + N_{\text{female}}}$$

This gender score for a character g_i can range from -1 (completely feminine) to 0 (gender-neutral) to 1 (completely masculine). The data contains a parts-per-million metric for each character during each of six birthing cohorts/generations. A average of gender specificity of a cohort (absolute value of gender score), weighted by parts-per-million in that cohort, could be used to see changes in gender-specificity over time. The birthing cohorts are: 1930-1959, 1960-1969, 1970-1979, 1980-1989, 1990-1999, 2000-2008.



Judging by the smooth trend line, gender-specificity remains relatively constant across birthing cohorts until a sharp drop towards gender-neutrality in the most recent cohort. The best fit linear trend line in blue does trend downwards, but this trend is not statistically significant, evidenced by the standard error band including possible trend lines with positive slopes. Alternatively, an ANOVA (analysis of variance) could be used to test whether there is a statistically significant difference between gender-specificity in different birthing cohorts. The hypotheses are:

$$H_0 : \text{Mean gender-specificity is the same across all cohorts}$$
$$H_A : \text{At least one cohort has a different mean gender-specificity}$$

The data should be weighted by parts-per-million in each cohort, which isn't exactly conducive to running a traditional ANOVA Instead, we can use the fact that a regression of one categorical feature is the same as a one-way ANOVA and fit a regression model weighted by parts-per-million. The model equation will be as follows…

$$y_i = \beta_0 + \sum_{k=1}^{6} \beta_k \times x_{ik} + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2 / w_{ik})$$

… where y_i represents the gender score for a character, and x_{ik} is an indicator the character is being weighted for Cohort k (each character will appear once in the dataset for each cohort). Each β_k is the difference in expected gender-specificity for Cohort k, and w_{ik} represents the parts-per-million for character i in birthing cohort k. An ANOVA can then be applied to the output of this model to get test the original null hypothesis that mean gender-specificity is the same across cohorts.

2.2 Subjective Name Characteristics By Gender

The data also provide three subjective metrics that attempt to quantify the sentiment of different characters and the characteristics of people who have them in their names. Here are the metric descriptions provided by the package author:

Name valence (positivity of character meaning): Ranges 1-5, *"Subjective ratings from 16 Chinese raters (9 males and 7 females; interrater reliability ICC = 0.921) on the positivity of all 2,614 given-name characters according to the meaning of each character (1 = strongly negative, 3 = neutral, 5 = strongly positive)."*

Name Warmth/Morality: Ranges 1-5, *"Subjective ratings from 10 Chinese raters (5 males and 5 females; interrater reliability ICC = 0.774) on how a person whose name contains each of the 2,614 given-name characters is likely to have warmth-related traits (1 = strongly unlikely to have, 3 = medium likelihood, 5 = strongly likely to have)."*

Name competence/assertiveness: Ranges 1-5, *"Subjective ratings from 10 Chinese raters (5 males and 5 females; interrater reliability ICC = 0.712) on how a person whose name contains each of the 2,614 given-name characters is likely to have competence-related traits (1 = strongly unlikely to have, 3 = medium likelihood, 5 = strongly likely to have)."*

3. Results

3.1 Anova Results

Coefficients Table

Predictor	Slope (Estimate)	p-value
Intercept	0.593	< 0.001
Time Period : 1960-1969	-0.003	0.683
Time Period :1970-1979	0.003	0.642
Time Period : 1980-1989	0.011	0.144
Time Period : 1990-1999	-0.004	0.597
Time Period : 2000-2008	-0.042	< 0.001

ANOVA Results

Independent Variable	F-statistic	P-value
Time Period	10.792	< 0.001

3.2 WLS Output

Valence

Decade	Estimated Coefficient ($\hat{\beta}_1$)	P-value
1930–1959	-0.053	< 0.001
1960–1969	-0.026	0.0436
1970–1979	-0.009	0.494
1980–1989	0.001	0.907
1990–1999	0.015	0.251
2000–2008	0.027	0.0394

Warmth

Decade	Estimated Coefficient ($\hat{\beta}_1$)	P-value
1930–1959	-0.054	< 0.001
1960–1969	-0.038	0.0007
1970–1979	-0.032	0.0035
1980–1989	-0.026	0.0132
1990–1999	-0.022	0.0359
2000–2008	-0.037	0.0010

Competence

Decade	Estimated Coefficient ($\hat{\beta}_1$)	P-value
1930–1959	0.088	< 0.001
1960–1969	0.134	< 0.001
1970–1979	0.172	< 0.001
1980–1989	0.202	< 0.001
1990–1999	0.204	< 0.001
2000–2008	0.206	< 0.001

4. Discussion

Chicken or the egg