STA440 Final Project

Matthew O'Donnell

2024-12-01

1. Introduction

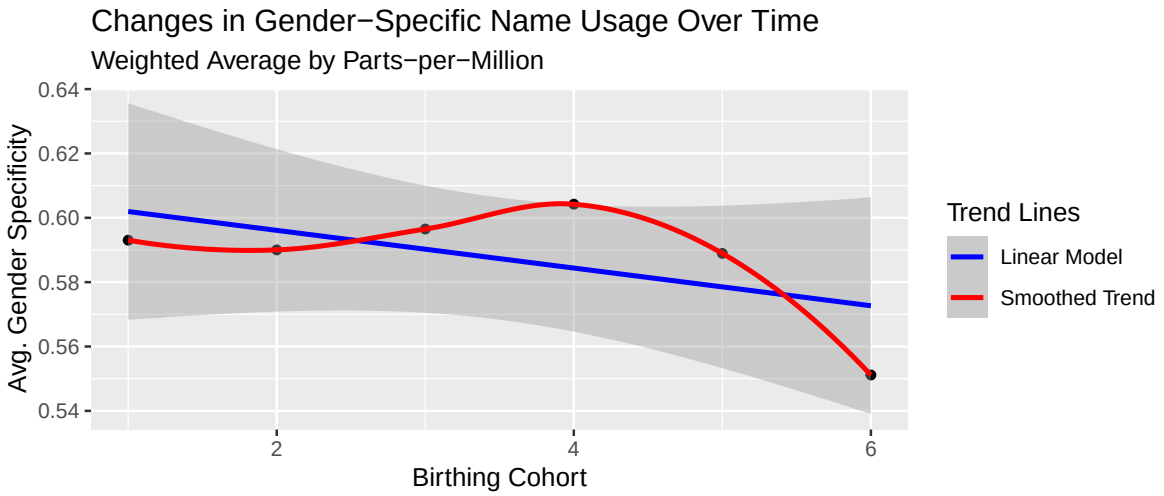
2. Methodology

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…

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

… where $y_i = |g_i|$ 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 between Cohort k and those in the 1939-1950 cohort (base case), 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. This model

makes the assumption that the gender-specificity, remains constant across birthing cohorts (ie the gender specificity of a character remains the same across different periods of time). This is not an ideal assumption to make, but it is required based on the limitations of the provided dataset. All hypothesis tests will be done at the $\alpha = 0.05$ level.

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)."*

For given characters, the relationship between gender score and the above subjective assessments can be assessed using weighted least squares. An example equation is as follows…

$$\begin{aligned} y_i &= \beta_0 + \beta_1 \times x_i + \epsilon_i \\ \epsilon_i &\sim N(0, \sigma^2 / w_{ik}) \end{aligned}$$

… where y_i represents a character's score in a given subjective metric, x_i represents the character's gender score, and β_1 represents the true coefficient for the subjective metric. This model can be fit separately for each birthing cohort, with w_{ik} representing parts-per-million for character i during cohort k. All hypothesis tests will be done at the $\alpha = 0.05$ level.

3. Results

3.1 Anova Results

Table 1 : Estimated WLS Coefficients By Time Period Effects on Gender-Specific Name Usage

Predictor	**Estimated Coefficient ($\hat{\beta}_k$)**	**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

Table 2 : ANOVA Metrics for Birthing Cohort (from WLS)

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

Looking at the estimated coefficients for each birthing cohort, it becomes clear that only the 2000-2008 period displays a statistically significant association with gender-specificity (which is negative, as would be expected based on the chart in 2.1). The results of the ANOVA test present a P-value < 0.001, which permits rejection of the null hypothesis that all cohorts have the same gender specificity.

3.2 Significance of Characteristic Effects

Table 3 : Valence Coefficients

Time Period	**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

Table 3 : Valence Coefficients		
2000–2008	0.027	0.0394

Table 4 : Warmth Coefficients		
Time Period	**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

Table 5 : Competence Coefficients		
Time Period	**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

The relationship between gender score and character valence only appears significant in the first and final birthing cohort. The former displays a negative association and the later a positive one. This would mean a negative association between subjective name valence and male-associated characters has developed into a positive one over time.

Coefficients for warmth and competence appear significant for each period, and the direction of each relationship remains constant across all periods. Table 4 shows associations between a female-associated characters and a warm perception, while Table 5 shows an association between male-associated characters and a competent perception.

4. Discussion

Chicken or the egg