Case Study #2

Il Shan Ng

Professor Laura Chihara
Math 315: Advanced Statistical Modeling

May 23 2016

1 Introduction

Literature censorship is a common occurrence, and books in particular get challenged by the public and subsequently banned for a variety of reasons. Previous studies have examined the reasons for why a book may be banned, but few have controlled for the surrounding social factors that may make such bans more or less likely to happen. My objective in this case study is to compare reasons for book bans after controlling for a few state-level demographic factors.

2 Exploratory Data Analysis

I use data on book challenges between January 2000 and November 2010, assembled by a team of students (Fast and Hegland, 2011) from the American Library Society. The data set consists of 1614 observations taken from challenges that occurred in 48 states. It also contains state-level demographic information obtained from the US Census Bureau and the Cook Political Report. One variable was missing in 11 observations, which I proceeded to remove from the data set.

The dependent variable of interest is `removed`, which indicates whether a book challenge was successful or not. Table 1 describes the numeric explanatory variables and some summary statistics. Table 2 identifies the categorical variables as well as the counts for each level (Yes and No). Other than `antifamily`, `occult` and `homosexuality`, the counts seem to be distributed well enough to allow for good analysis. To avoid model-fitting issues, I center and scale `cmedin` and `days2000`.

Table 1: Summary statistics for numeric variables

Variable	Description	Min	Max	Mean	SD
`pvi2`	Political Value Index; positive values indicate a more Democratic leaning	-20.2	13.4	-4.25	7.16
`cphers`	percentage of high school graduates in a state	-6.66	8.74	-0.0339	4.25
`cmedin`	median state income	-8466	19940	2595	4727
`cperba`	percentage of college graduates in a state	-9.22	9.18	-0.0042	3.08
`days2000`	date of challenge; days after Jan 1 2000	0	3904	2036	1176.5

Table 2: Counts for categorical variables

Variable	Description	Yes	No
`obama`	whether a challenge was made when Obama was President	447	1156
`freqchal`	whether the book's author was frequently challenged	269	1345
`sexexp`	book was challenge for sexually explicit material	571	1043
`antifamily`	challenged for anti-family material	46	1568
`occult`	challenged for material about the occult	96	1518
`language`	challenged for inappropriate language	519	1095
`homosexuality`	challenged for material about homosexuality	112	1502
`violence`	challenged for violent material	231	1383

Other information that will not be used as explanatory variables include the `book` ID number, `author` and `state` in which the challenge occurred. These will instead be used as random effects to respect the multi-level nature of the data. Because the same book may have been challenged in different states, the `book` and `state` random effects are crossed.

One variable that is likely to influence book bans is the political leaning of the state in which the book was banned. Figure 1 shows the distribution of `pvi2` grouped by successful and unsuccessful challenges, and suggests that bans tended to occur more in states that had a Republican leaning (negative `pvi2`). Figure 2 shows that whether or not the author of a book was frequently challenged had little effect on the probability of a book getting banned, and so we expect `freqchal` to not be in the final model.

Figure 1: Political value index grouped by whether or not challenge was successful

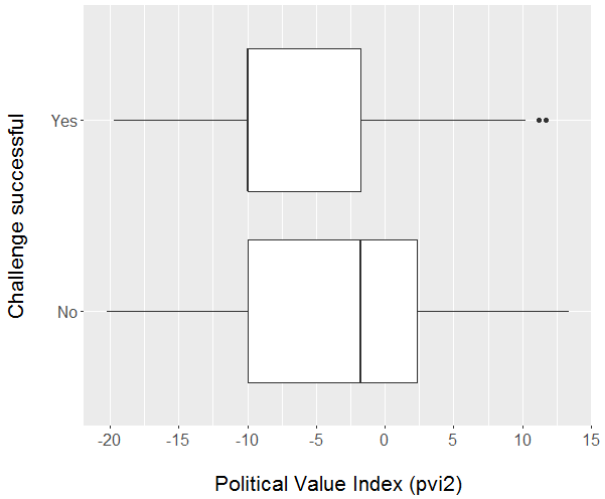
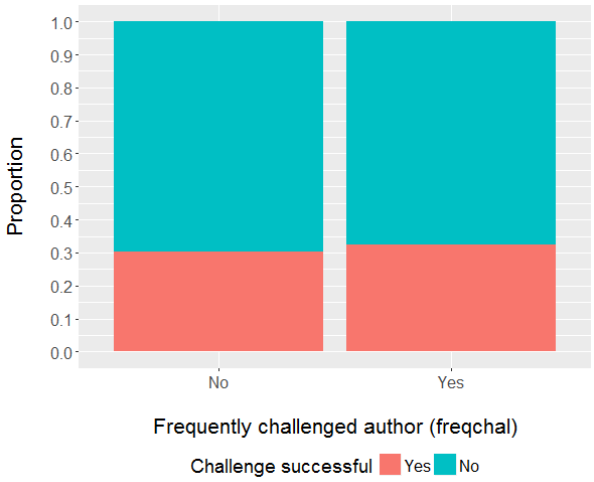


Figure 2: Proportion of successful challenges grouped by whether or not author was frequently challenged



3 Results

Since the dependent variable is binary (Yes/No), I modeled the probability of a successful book challenge using binary logistic regression. Let π_{ijk} represent the probability that the k^{th} challenge for the i^{th} book in the j^{th} state succeeds. The final hierarchical model is shown below.

Level One (book i state j challenge k):

$$\ln\left(\frac{\pi_{ijk}}{1-\pi_{ijk}}\right) = a_{ij} + \beta_1 \text{days2000}_{ijk} + \beta_2 \text{obama}_{ijk} + \beta_3 \text{sexexp}_{ijk} + \beta_4 \text{antifamily}_{ijk} + \beta_5 \text{language}_{ijk}$$

Level Two (book i state j):

$$a_{ij} = \alpha_0 + \alpha_1 \text{pvi2}_j + \alpha_2 \text{cperhs}_j + u_i + v_j$$

where u_i is the random effect for book i with $u_i \sim N(0, \sigma_u)$, and v_i is the random effect for state j with $v_j \sim N(0, \sigma_v)$. As expected, `freqchal` ended up being insignificant. The random effect for `author` also caused serious model-fitting issues and so was removed to achieve stability. I also found all two-way interaction terms involving the variables in the final model above to be unnecessary. Tables 3 and 4 give the estimates for the fixed effect parameters and variance components. `pvi2` and `cperhs` are significant at the 10% level, while all other fixed effects are significant at the 5% level.

Table 3: Estimates for fixed effect parameters

Coefficient	Notation	Estimate	Standard error	z-value	p value
Intercept	α_0	-1.052	0.217	-4.853	0.00000121 ***
`pvi2`	α_1	-0.0394	0.0201	-1.955	0.0506 .
`cperhs`	α_2	-0.074	0.0397	-1.865	0.0622 .
`days2000`	β_1	0.236	0.0971	2.429	0.0152 *
`obama`	β_2	-0.531	0.218	-2.429	0.0151 *
`sexexp`	β_3	0.238	0.131	2.044	0.0409 *
`antifamily`	β_4	-0.999	0.455	-2.20	0.0278 *
`language`	β_5	0.375	0.134	2.805	0.00503 **

Table 4: Estimates for variance components

Component	Notation	Estimate
SD of `book` random intercept	σ_u	0.5322
SD of `state` slope	σ_v	0.7953

In refining my model, I removed 31 extreme outliers that were flagged in a plot of residuals against fitted values. I also removed book 868 (Harry Potter series), which had an unusually small random intercept. Doing so removed another 46 observations and so in total, about 5.5% of the original challenges was not used in this analysis.

4 Discussion

From Table 3, we see that the estimated coefficient for `pvi2` is negative. This implies that a greater Democratic leaning is associated with smaller odds of a book challenge being successful, confirming my earlier exploratory data analysis. The negative coefficient on `cperhs` also indicates that a state having higher high school graduation rates is associated with smaller odds of a book challenge being successful.

Among the six reasons for book challenges identified earlier, only three remain in the final model. Challenges that were made on the basis of sexually explicit material or inappropriate language tended to succeed more. In particular, after controlling for state demographics and the time of challenge, having sexually explicit material as the reason is associated with a 27% increase in the odds of success relative to having no reason, and having inappropriate language as the reason is associated with a 45% increase in the odds of success relative to having no reason. On the other hand, having anti-family material as the reason is associated with a 63% decrease in the odds of success relative to having no reason, a finding which may be somewhat surprising.

In conclusion, I have found that after controlling for certain state-level factors, using inappropriate language as the reason for challenge was most likely to lead to a successful challenge. Using sexually explicit material as the reason also improved the odds of a successful challenge, while using anti-family material as the reason decreased the odds of a successful challenge. Using material about the occult, material about homosexuality, and violent materials as reasons did not significantly affect the odds of a book getting banned.