Patterns and Predictability in Borrower Behavior

Alex Leslie

Rutgers University English

@azleslie — azleslie.com

April 10, 2021

Latent Variable Interpretation of Discrete Choice

An individual i makes choice c over all alternative options a if and only if they derive greater utility U from it.

$$U_{ic} > U_{ia} \, \forall \, a \neq c \tag{1}$$

In the case of borrowing habits, I propose interpreting this latent variable as taste.

We don't know the true taste U; we have independent variables that allow us to measure a representative taste V, and we assume a distribution of error ϵ that covers the difference.

$$U_{ic} = V_{ic} + \epsilon_{ic} \tag{2}$$

Y_i is the outcome, either 1 or 0, for an individual observed choice that manifests U: a borrower's checkout that manifests taste. The goal of a logistic model is to estimate the probability p_i of this outcome.

Logistic Regression

The representative taste V portion of Equation 2 looks like linear regression.

$$V_{ic} = \beta_0 + \beta_1 x_{1ic} + \beta_2 x_{2ic} + ... + \beta_k x_{kic} \tag{3}$$

x_{1ic} is the first independent variable for an individual observation making a choice and so on until the last kth independent variable; β_1 is the coefficient that the model assigns to that independent variable. It does so via an iterative Maximum Likelihood Estimation; this occurs in the model's training phase. See Function

Latent Variable Interpretation Continued

What does the distribution ϵ in Equation 2 above correspond to? The interpretation of this element impacts our interpretation of what probability refers to in the model. These are some possible avenues.

- Describes portion of population
- Describes our confidence
- Describes individual idiosyncrasy

In logistic regression, this ends up being equivalent to logistic distribution, where p_{ic} is the probability of an individual observation making a choice.

$$\textit{logit link function} = \log_e \left(\frac{p_{ic}}{1 - p_{ic}} \right) \tag{4}$$

The Curve

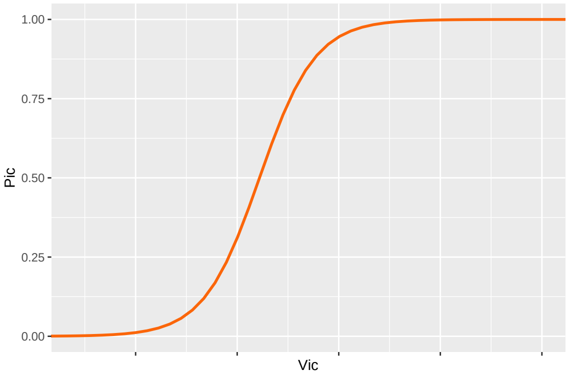


Figure 1: Graph of a characteristic logit curve.

Logistic Regression Continued

Put together, the pieces look like so:

$$\log_e\left(\frac{p_{ic}}{1-p_{ic}}\right) = \beta_0 + \beta_1 x_{1ic} + \beta_2 x_{2ic} + \ldots + \beta_k x_{kic} \tag{5}$$

Solving for p (and condensing notation for independent variables and coefficients) results in:

$$P_{ic} = \frac{e^{\beta' x_{ic}}}{1 + e^{\beta' x_{ic}}} \tag{6}$$

In the case of more than two choices, the denominator becomes $\sum_a e^{\beta' x_{ia}}$, the sum of the exponents for each possible choice.

Assumptions

- Choices must be:
  - Exclusive (an individual can only choose one)
  - Exhaustive (represent all available choices)
  - Finite
- Minimal or no multicollinearity among independent variables
- Observations must be independent from one another
- One in ten rule? (independent variables to positive events)

General Muncie Library Stats

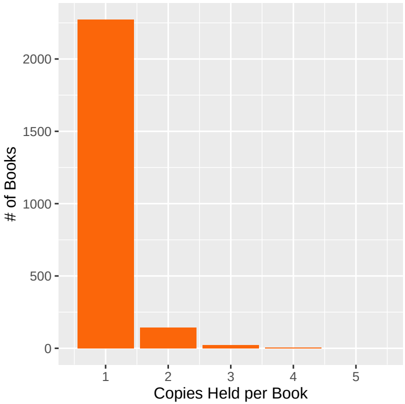


Figure 2: Copies held per book.

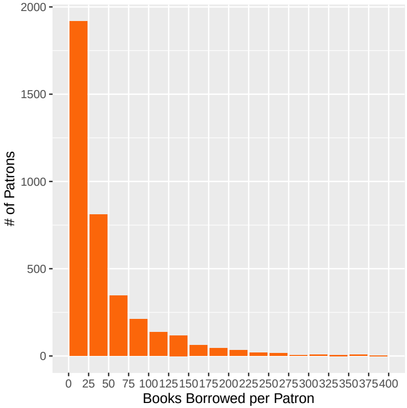


Figure 3: Books borrowed per patron.

A Preposterous Graph

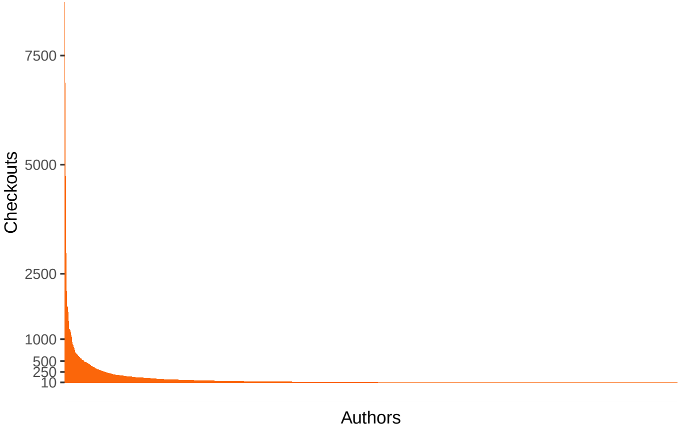


Figure 4: Total number of checkouts for books by each author held in the Muncie Public Library.

Checkout Tails

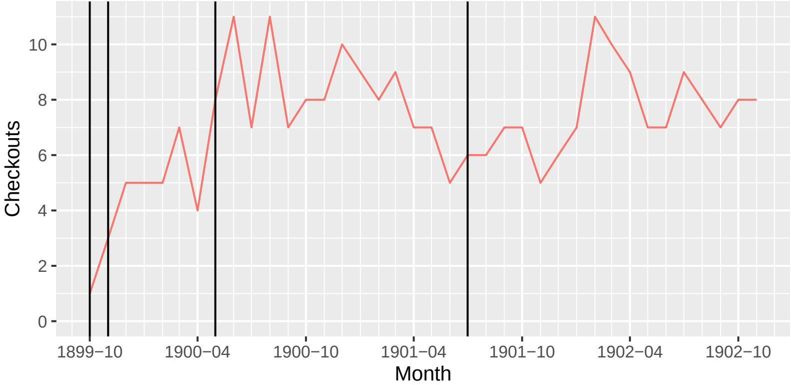


Figure 5: Muncie Public Library checkouts of Booth Tarkington's *The Gentleman from Indiana* (1899). Orange line marks checkouts per month, black vertical lines mark the accession of a new copy.

Checkout Tails 2

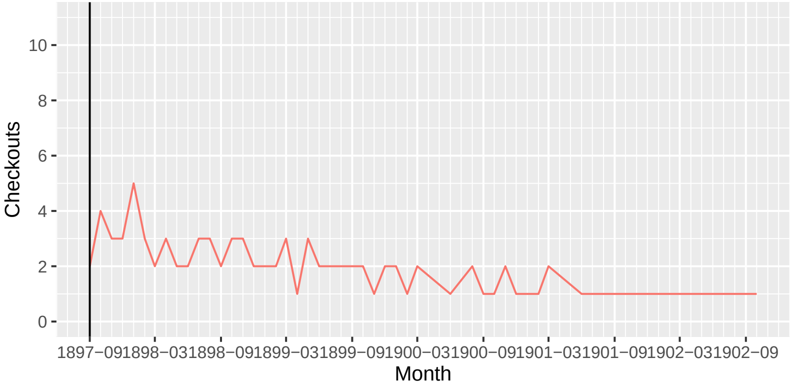


Figure 6: Muncie Public Library checkouts of William Dean Howells' *The Landlord at Lion's Head* (1897). Orange line marks checkouts per month, black vertical lines mark the accession of a new copy.

Back to Assumptions

- Choices must be:
  - Exclusive (an individual can only choose one)
  - Exhaustive (represent all available choices)
  - Finite
- Minimal or no multicollinearity among independent variables
- Observations must be independent from one another
- One in ten rule? (independent variables to positive events)

Classifying and Assessing Results

		True Condition	
		Positive (Borrowed)	Negative (Unborrowed)
Predicted Condition	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Table 1: Classification table.

Basic measures of a model's fit:

- Accuracy, the proportion of correct predictions: (TP+TN) / (TP+TN+FP+FN)
- Sensitivity, the proportion of correctly predicted positive cases: TP / (TP+FN)
- Specificity, the proportion of correctly predicted non-events: TN / (TN+FP)

Multiple Models

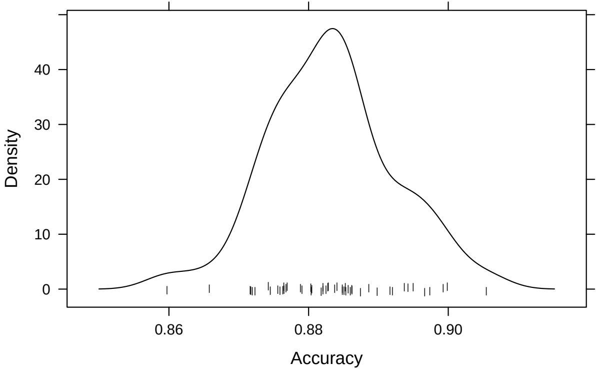


Figure 7: The accuracies of 50 logit models estimating the probability that Muncie Public Library patrons borrowed a novel by Mary Johnston.

The Class Imbalance Menace

Running models on checkout data as-is, runs into class imbalance issues. Because checkouts are, comparatively, rare events, models give them comparatively less weight.

Here are results for Mary Johnston, author of the runaway bestseller *To Have and to Hold* (1899), in a model of a subset defined by patrons who borrowed books by > 5 authors and authors who had > 200 total checkouts.

	Actually Borrowed	Actually Unborrowed
Predicted Borrowed	15	16
Predicted Unborrowed	42	499

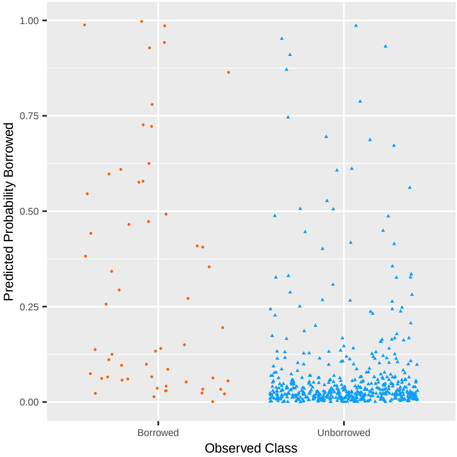
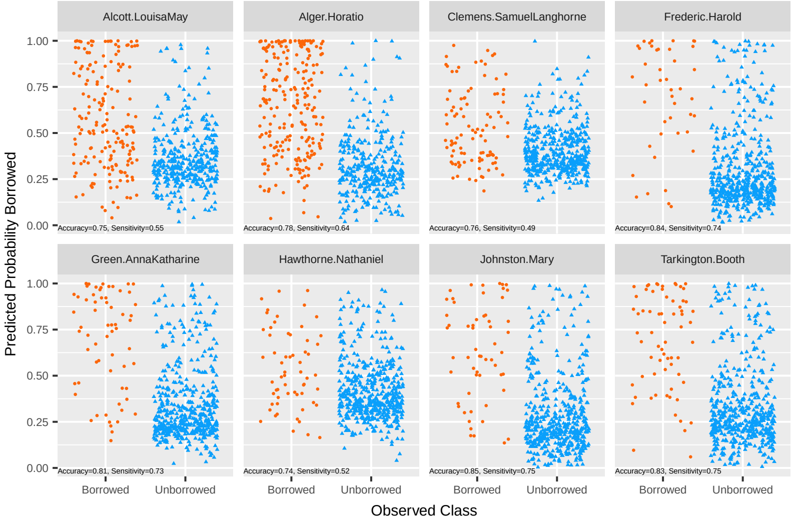


Figure 8: Results for Mary Johnston.

Model Parameters Compared

Patron Author Minimum	Total Author Minimum	Sampling Method	Accuracy	Sensitivity	# of Authors	# of Patrons
5	200	normal	0.8766	0.3320	122	2862
5	200	down	0.7052	0.6457	122	2862
5	300	down	0.7511	0.6564	77	2853
5	400	down	0.7756	0.6453	39	2804
5	200	rose	0.8086	0.6108	122	2862
5	300	rose	0.7994	0.6232	77	2853
5	400	rose	0.7946	0.6175	39	2804
10	150	down	0.6509	0.6406	154	2294
10	200	down	0.6887	0.6535	117	2294
10	300	down	0.7264	0.6455	75	2293
20	150	down	0.6381	0.6361	138	1524
10	150	rose	0.7917	0.6090	154	2294
10	200	rose	0.7809	0.6143	117	2294
10	300	rose	0.7736	0.6170	75	2293
20	150	rose	0.7463	0.6050	138	1524
60	100	normal	0.6557	0.5451	98	413
60	100	rose	0.5911	0.5092	98	413
60	80	normal	0.6355	0.5275	136	413
60	80	rose	0.5998	0.5188	136	413
80	60	normal	0.5713	0.5428	117	219
80	80	normal	0.6030	0.5893	69	219

Model Results



Diverging Fit between Authors

Key factors that improve the fit of models of whether patrons borrowed an author's work:

- Authors who were more popular as measured by more total checkouts – but not extremely popular;
- Authors with more books held by the library (which partly though not entirely impacts total checkouts and is partly though not entirely impacted by the total number of books that author wrote);
- Authors who wrote in a particular genre, movement, or otherwise relatively niche part of the literary marketplace;
- Authors who did not belong to the emergent popular canon.

Maximum Likelihood

This function is repeated, each time picking a new vector of coefficients β (one for each independent variable) to find that which maximizes the value (recalling the equation for P_{ic} on the previous slide)

$$LL(\beta) = \sum_{i} \sum_{c} y_{ic} \log_e(P_{ic}) \tag{7}$$

Return