MSE 226: Project Part 2 - Inference and Causality

2. Fitting on Test Data.

The test set gave an RMSE of **0.683**, which is higher than the cross-validated RMSE (**0.657**). As the generalization error is lower than the error on the test set, this may mean that the model was slightly overfit on the training data.

3a. For your chosen model, report which coefficients are statistically significant in the regression output.

We used a log transformation of visits because the raw visitation counts exhibit strong right-skew and heteroskedasticity: variance increases substantially with other covariates. Transforming the outcome leads to lower-variance coefficient estimates and clearer, more interpretable inference. We also applied log transformations to several covariates for related reasons: the covariates are heavily right-skewed and logging them linearizes the associations (**See Table 1**). Many of the variables are statistically significant at the 95% confidence level. The significant covariates include log(LSA Population), log1p(Print Volumes), log1p(Ebook Volumes), log(County Population), Number of Bookmobiles, Having an Overdue Policy, Not Being a Member of a Federation or Cooperative, being in a Rural Region, and being located in the Great Lakes, New England, the Rocky Mountains, or the Southeast (**See Table 2**). Statistical significance at this level means that if the true coefficient were zero, then we would expect fewer than 5% of repeated samples to produce a coefficient as large in magnitude as the one we observe, given that the model assumptions hold. For example, the estimated coefficient on log(LSA Population) is approximately **0.42** and statistically significant. This implies that a 1% increase in LSA population is associated with roughly a **0.42%** increase in visits, provided all other covariates are held constant.

3b. Now fit your chosen model on the held-out test data, and look at whether the same coefficients are significant.

When we fit the same model on the held-out test data (50-50 split), nearly all of the relationships that were statistically significant at a 95% confidence level in the training set remained significant (**See Table 2**). The only exception was the indicator for "beac_codeNew England." No variables that were non-significant in the training sample became significant in the test sample. In the training data, the coefficient on "beac_codeNew England" was **–0.152**, with a 95% confidence interval of **[–0.275, –0.030]**, narrowly excluding 0. In the test data, the estimated coefficient was slightly smaller in magnitude (**–0.109**) with a confidence interval of **[–0.241, 0.024]**, barely including 0. This modest shift is consistent with sampling variability arising from the dataset split (**344** units in the training set and **327** in the test set), and may indicate that the effect of being in New England is not practically significant.

3c. Use the bootstrap to estimate confidence intervals for each of your regression coefficients.

Comparing the standard regression confidence intervals to the bootstrap confidence intervals reveals a strong consistency. Only one confidence interval switched from not including zero in the original inference to including zero when creating bootstrap confidence intervals. This was the confidence interval for whether a library is not a member of a federation or cooperative. In the original inference, the 95% confidence interval for this variable is **[-0.115, -0.00133]**, extremely close to including 0, and in the bootstrap, the 95% confidence interval is **[-0.122, 0.00278]**, slightly wider, and includes 0. The bootstrap confidence intervals were wider for the majority of coefficients, besides "num_bookmobiles",

"num_lib_branches", "locale_codeRural", "locale_codeSuburban", and "locale_codeTown", where they were smaller. The original and the bootstrap standard errors are all quite similar.

We would report the bootstrap confidence intervals to a stakeholder. This method results in more robust results since it does not rely on strict distributional assumptions. By repeatedly resampling the data, the bootstrap estimates the sampling distribution of the coefficients, providing a more reliable measure of the parameter's stability. In practice, the bootstrap is more conservative, helping the stakeholder avoid basing policy on a fragile finding. This ensures that operational decisions are guided by effects that are consistently and reliably supported by the data.

3d. Discuss multiple hypothesis testing in your analysis.

Since there are **20** hypothesis tests (variables and binarized categoricals) at a significance level of 5%, and each test has a 5% Type I error rate under a true null, the probability of obtaining at least one false positive, and thus too many seemingly significant predictors, increases with the number of tests. We can apply the Bonferroni correction, which guarantees that declaring at least one false positive among all hypotheses is at most 5% (instead of guaranteeing at most 5% Type I error rate for each of the 20 individual hypothesis tests). It achieves this by using a per-test significance level, denoted as alpha/m. In our case, alpha is 0.05, and m is 20, so the per-test significance level is **0.0025** (or, in practice, multiply the p-value by 20). We find that the Bonferroni correction renders the coefficient on the library not being a member of a federation or cooperative not significant, along with all location-based coefficients. The Benjamini-Hochberg procedure is less conservative and controls the false discovery rate, i.e., the expected proportion of rejected hypotheses that are actually null. Running the Benjamini-Hochberg procedure only changed the significance of the coefficient on whether the library was not a member of a federation or cooperative, in line with the bootstrap and the Bonferroni correction. (**See Table 3).**

The Bonferroni correction is too conservative for our application, while the Benjamini-Hochberg procedure gives the same significant variables as the bootstrap. Therefore, we would not discuss multiple testing correction when presenting our analysis to Julie Balutis, since this may overcomplicate the analysis without changing the results.

3e. Explain why your model-building and inference process suffers from post-selection inference. (If you don't think it does, defend why in this part.)

We did not choose our variables based on statistical significance; the list of predictors was determined at the outset based on subject-matter considerations and prior research. However, our choice of transformations (logging skewed variables to address heteroskedasticity) was informed by examining the data. Because these modeling decisions were made after inspecting distributions and residual patterns, our inference involves post-selection: standard errors and p-values do not fully account for the fact that the model specification was refined using the data itself. That said, the variable list was not selected by significance testing or automated procedures, and we did not add or remove predictors in response to their p-values.

4a. Revisit the stakeholder scenario you identified in Part 1. How do your statistical inference results inform the specific decisions your stakeholder needs to make?

Our inference results give Julie Balutis concrete guidance on which library characteristics are meaningfully associated with visitation and, therefore, may inform IMLS grant decisions. A practically significant association is with LSA population size, which remains significant across the training set, test set, and bootstrap. This suggests that Julie can use the population as a reliable baseline indicator of expected visitations when comparing similar applicants. Two additional variables show practically meaningful associations: systems with more branches are associated with about **1.6%** fewer visits per additional branch, and libraries with overdue policies show about **12%** lower visitation. While these relationships are not causal, they can still help Julie interpret differences in visitations across library systems and understand which operational features tend to coincide with higher or lower visitation. In contrast, geographic patterns may be more inconsistent and are correlated with population differences, so they should not be treated as decisive signals in funding decisions. Overall, the inference results highlight which associations may inform Julie's interpretation of proposals and which patterns are more uncertain.

4b. Identify one statistically significant relationship from your analysis where your stakeholder would particularly care about whether the relationship is causal rather than merely correlational.

A statistically significant relationship that our stakeholders would care about being causal is the effect of having an overdue policy (versus none) on library visits. If eliminating overdue policies *causes* an increase in library visits, the IMLS could confidently create grant incentives to encourage all libraries to adopt overdue policies. This would be a simple intervention to deliver the measurable benefit of increased library use, aligning with the agency's mission to ensure efficient use of Federal funds.

If the relationship is just correlation, it means that there is no definitive evidence that changing the overdue policy would have an impact on visitation. This may be a result of unobserved factors (a modern administrative approach, higher local funding, a specific community demographic) that are the true drivers of higher visits. In this scenario, encouraging libraries to change their overdue policy may be a poor use of IMLS influence. For a binary policy like this, knowing whether the effect is causal directly shapes what policies the office recommends and funds.

4c. For the relationship you identified in part (b), evaluate whether the evidence supports a causal interpretation.

We estimate the treatment effect using a cross-fitted AIPW estimator. A 5-fold Random Forest classifier generates out-of-sample propensity scores for receiving the overdue-policy treatment, while separate Random Forest models predict potential outcomes under both treatment and control conditions. These cross-fitted components are combined to yield a Cross-Fit average treatment effect of **-0.0933 [-0.1605, -0.0262]**, indicating lower log visitation under the policy. While the confidence interval does not include 0, the average treatment effect should be brought into question due to confounding, selection effects, and measurement issues, as well as the proximity of the confidence interval to 0.

Particularly, AIPW cannot account for unobserved confounding. State and local financing is one such example. Those libraries without overdue policies may be located in communities with higher state funding, allowing the financial freedom to forgo overdue policies. This higher funding could be the true

cause of higher visits (potentially funding more programs, hours, or better staff). Since local financing is unobserved in our model, the estimated negative effect of the overdue policy may be spurious.

Additionally, the decision to have an overdue policy may not be random. Libraries that maintain such policies may be those that are historically more conservative, financially stressed, or serve communities with different expectations for public services. These unobserved differences are selection effects that may drive both the policy and the outcome. Additionally, our data is self-reported, and we removed libraries with incomplete data. There may be systematic differences between libraries that report fully and those that do not. This potentially non-random selection of libraries in our dataset biases the association.

Finally, our treatment variable is a binary indicator. This does not account for the severity, amount, or enforcement of the overdue policy. Different types of overdue policies may have different effects on the visitation. Another issue is that the data does not indicate when the overdue policy was implemented. If implemented mid-year, this may change the annual visits, without us knowing. These are measurement issues. Based on the findings and limitations, particularly in potential unconfoundedness, we would not recommend that IMLS implement a grant-seeking process that emphasizes removing overdue policies.

4d. What additional data would most strengthen your ability to make credible causal claims about the relationship in part (b)?

The strongest evidence for a credible causal claim would come from a Randomized Controlled Trial in which libraries with overdue policies are randomly assigned either to keep their current policy or to eliminate it. Random assignment would eliminate both observed and unobserved confounding, ensuring that differences in annual visitation can be attributed to the policy itself. This design would require (1) a large, representative sample of libraries and (2) true random assignment of the policy change. We would also want to observe key covariates such as local funding, community socioeconomic characteristics, and library financial stress, factors that may currently confound our observational estimates and help interpret and generalize the results.

However, implementing such an RCT may be challenging. Some libraries exist in larger municipal systems, which guide policy. Making a change of this kind may require system-wide approval and cannot be easily randomized at the branch level. Branch-level randomization may also increase administrative burden and conflict with expectations of uniform service. Ethical concerns also arise: deliberately assigning some branches to policies believed to be beneficial, or withholding them, may not be acceptable to staff, patrons, or governing boards. Finally, spillover is a threat: patrons could shift their borrowing to nearby branches with more favorable policies, contaminating treatment assignment and biasing the estimated effects.

5a. We used ChatGPT, Gemini, and Perplexity

5b. AI was particularly helpful for basic code generation and spell-checking, but it performed poorly when interpreting statistical findings in the specific, rigorous manner required for this class. The models often blurred the line between correlation and causation, especially when producing recommendations, which made their interpretations unreliable without careful human correction.

MSE 226: Project Part 2 - Inference and Causality

Appendix

Table 1.

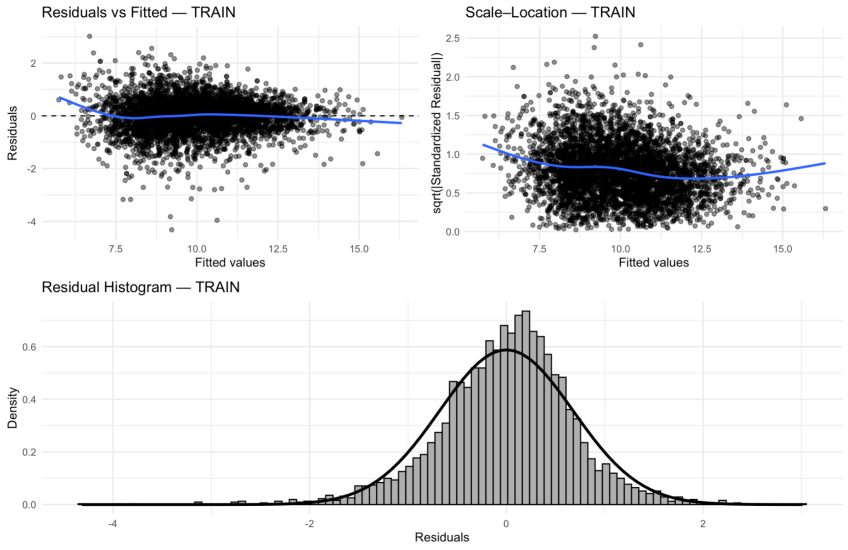

Table 2.

estimate_train	term	train_p	significant_train	estimate_test	test_p	significant_test
<dbl>	<chr>	<dbl>	<lgl>	<dbl>	<dbl>	<lgl>
-2.08	(Intercept)	1.08e- 15	TRUE	-1.38	4.15e- 8	TRUE
0.421	log(population_lsa)	1.58e-147	TRUE	0.448	4.11e-181	TRUE
0.703	log1p(print_volumes)	1.06e-199	TRUE	0.601	2.67e-203	TRUE
0.0578	log1p(ebook_volumes)	1.60e- 48	TRUE	0.0773	2.46e- 84	TRUE
0.0456	log(county_population)	5.00e- 6	TRUE	0.0317	1.73e- 3	TRUE
-0.0141	num_bookmobiles	6.57e- 1	FALSE	0.0212	5.39e- 1	FALSE
-0.0158	num_lib_branches	5.88e- 6	TRUE	-0.0119	1.25e- 3	TRUE
-0.124	overdue_policyHas Overdue Pol…	2.77e- 8	TRUE	-0.119	1.99e- 7	TRUE
-0.0583	interlibrary_relation_codeNot…	4.49e- 2	TRUE	-0.0875	4.05e- 3	TRUE
0.111	fscs_definition_codeMeets FSC…	4.58e- 1	FALSE	0.253	1.00e- 1	FALSE
-0.179	locale_codeRural	3.14e- 3	TRUE	-0.264	1.84e- 5	TRUE
0.0809	locale_codeSuburb	1.11e- 1	FALSE	0.0386	4.55e- 1	FALSE
0.0192	locale_codeTown	7.21e- 1	FALSE	-0.0817	1.34e- 1	FALSE
-0.160	beac_codeGreat Lakes	3.90e- 3	TRUE	-0.157	8.71e- 3	TRUE
-0.0427	beac_codeMid East	4.47e- 1	FALSE	-0.0510	4.02e- 1	FALSE
-0.152	beac_codeNew England	1.47e- 2	TRUE	-0.109	1.09e- 1	FALSE
-0.0263	beac_codePlains	6.52e- 1	FALSE	-0.0769	2.16e- 1	FALSE
0.197	beac_codeRocky Mountains	5.41e- 3	TRUE	0.215	5.52e- 3	TRUE
-0.311	beac_codeSoutheast	2.91e- 7	TRUE	-0.241	1.82e- 4	TRUE
-0.111	beac_codeSouthwest	6.58e- 2	FALSE	-0.0804	2.16e- 1	FALSE

Table 3.

term	estimate	std_error	t_value	pr_t	significant_train	p_raw	p_bonf	p_bh	sig_raw	sig_bonf	sig_bh
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<lgl>	<dbl>	<dbl>	<dbl>	<lgl>	<lgl>	<lgl>
(Intercept)	-2.08	0.259	-8.05	1.08e- 15	TRUE	1.08e- 15	2.17e- 14	5.42e- 15	TRUE	TRUE	TRUE
log(population_lsa)	0.421	0.0156	26.9	1.58e-147	TRUE	1.58e-147	3.17e-146	1.58e-146	TRUE	TRUE	TRUE
log1p(print_volumes)	0.703	0.0220	31.9	1.06e-199	TRUE	1.06e-199	2.12e-198	2.12e-198	TRUE	TRUE	TRUE
log1p(ebook_volumes)	0.0578	0.00390	14.8	1.60e- 48	TRUE	1.60e- 48	3.21e- 47	1.07e- 47	TRUE	TRUE	TRUE
log(county_population)	0.0456	0.00999	4.57	5.00e- 6	TRUE	5.00e- 6	9.99e- 5	1.43e- 5	TRUE	TRUE	TRUE
num_bookmobiles	-0.0141	0.0318	-0.445	6.57e- 1	FALSE	6.57e- 1	1 e+ 0	6.91e- 1	FALSE	FALSE	FALSE
num_lib_branches	-0.0158	0.00348	-4.54	5.88e- 6	TRUE	5.88e- 6	1.18e- 4	1.47e- 5	TRUE	TRUE	TRUE
overdue_policyHas Overdue Policy	-0.124	0.0223	-5.57	2.77e- 8	TRUE	2.77e- 8	5.54e- 7	1.11e- 7	TRUE	TRUE	TRUE
interlibrary_relation_codeNot a memb…	-0.0583	0.0291	-2.01	4.49e- 2	TRUE	4.49e- 2	8.98e- 1	6.91e- 2	TRUE	FALSE	FALSE
fscs_definition_codeMeets FSCS Publi…	0.111	0.149	0.742	4.58e- 1	FALSE	4.58e- 1	1 e+ 0	5.39e- 1	FALSE	FALSE	FALSE
locale_codeRural	-0.179	0.0606	-2.96	3.14e- 3	TRUE	3.14e- 3	6.27e- 2	6.97e- 3	TRUE	FALSE	TRUE
locale_codeSuburb	0.0809	0.0507	1.59	1.11e- 1	FALSE	1.11e- 1	1 e+ 0	1.48e- 1	FALSE	FALSE	FALSE
locale_codeTown	0.0192	0.0535	0.358	7.21e- 1	FALSE	7.21e- 1	1 e+ 0	7.21e- 1	FALSE	FALSE	FALSE
beac_codeGreat Lakes	-0.160	0.0552	-2.89	3.90e- 3	TRUE	3.90e- 3	7.80e- 2	7.80e- 3	TRUE	FALSE	TRUE
beac_codeMid East	-0.0427	0.0561	-0.761	4.47e- 1	FALSE	4.47e- 1	1 e+ 0	5.39e- 1	FALSE	FALSE	FALSE
beac_codeNew England	-0.152	0.0625	-2.44	1.47e- 2	TRUE	1.47e- 2	2.94e- 1	2.45e- 2	TRUE	FALSE	TRUE
beac_codePlains	-0.0263	0.0584	-0.451	6.52e- 1	FALSE	6.52e- 1	1 e+ 0	6.91e- 1	FALSE	FALSE	FALSE
beac_codeRocky Mountains	0.197	0.0708	2.78	5.41e- 3	TRUE	5.41e- 3	1.08e- 1	9.84e- 3	TRUE	FALSE	TRUE
beac_codeSoutheast	-0.311	0.0606	-5.14	2.91e- 7	TRUE	2.91e- 7	5.83e- 6	9.72e- 7	TRUE	TRUE	TRUE
beac_codeSouthwest	-0.111	0.0606	-1.84	6.58e- 2	FALSE	6.58e- 2	1 e+ 0	9.40e- 2	FALSE	FALSE	FALSE