Understanding the Relationship between Attributes and Whether DIH Exists*

A Data Analysis of Facilities of Toronto Libraries

Yanzun Jiang

November 26, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

Table of contents

*Code and data are available at: https://github.com/Stary54264/relationship_between_attributes_and_whether_dih_exists_in_toronto_libraries.

1 Introduction

Despite the growing importance of innovation hubs in the global economy, the factors influencing the presence of Designated Innovation Hubs (DIH) in facilities remain underexplored and needed to be filled. This paper seeks to identify the key factors that influence whether a library in Toronto has a DIH, focusing on attributes such as area size, the number of public parking spots, the number of workstations in the library, and the year the library was built.

To investigate these factors, a logistic regression model was utilized, which is ideal for predicting binary outcomes (the presence or absence of a DIH). This analysis aims to estimate the likelihood of a facility having a DIH based on these attributes. The dataset used in this study contains information on various libraries in Toronto. After data cleaning, relevant variables was selected to do data analysis. The logistic regression model will reveal how each attribute (area size, parking availability, number of workstations, and age of the building) impacts the presence of a DIH.

The estimand in this paper is the presence of a Designated Innovation Hub (DIH) in a library. However, accurately observing which libraries have a DIH is not straightforward due to several challenges faced by urban planners. They may not have access to complete or up-to-date information about all libraries. This lack of comprehensive data, combined with regional variations and different reporting standards, makes it difficult to directly know the presence of a DIH across all libraries. To address this issue, a logistic regression model was built, since it allows us to estimate the probability of a facility having a DIH based on key attributes available in the dataset. By doing so, we aim to provide urban planners with insights into which factors most strongly influence the presence of DIHs, despite the challenges in directly measuring this estimand.

Results

DIHs are vital in fostering entrepreneurship, driving economic development, and supporting technological advancements. Understanding the attributes that contribute to the likelihood of a library of having a DIH could provide valuable insights for urban planners and policymakers.

The remainder of this paper is structured as follows: Section 2 discusses the data used in this analysis, including an overview of the dataset and key descriptive statistics. Section 3 outlines the logistic regression model, its assumptions, and its interpretation. Section 4 presents the results of the model, including the significance of each attribute in predicting the presence of a DIH. Section 5 offers a discussion of the findings, examines how these attributes affect DIH presence, and lists some weakness of this analysis. Statistical analysis is performed using `R` (R Core Team 2023), with packages `tidyverse` (Wickham et al. 2019), `arrow` (Richardson et al. 2024), `janitor` (Firke 2023), `testthat` (Wickham 2011), `rstanarm` (Brilleman et al. 2018), `here` (Müller 2020), and `knitr` (Xie 2014).

2 Data

2.1 Overview

Our data (Toronto Public Library 2024)…. Following Alexander (2023), we consider…

Overview text

2.2 Measurement

Some paragraphs about how we go from a phenomena in the world to an entry in the dataset.

2.3 Outcome variables

Add graphs, tables and text. Use sub-sub-headings for each outcome variable or update the subheading to be singular.

2.4 Predictor variables

Add graphs, tables and text.

Use sub-sub-headings for each outcome variable and feel free to combine a few into one if they go together naturally.

3 Model

The goal of our modelling strategy is twofold. Firstly,…

Here we briefly describe the Bayesian analysis model used to investigate… Background details and diagnostics are included in Appendix B.

3.1 Model set-up

Define y_i as the number of seconds that the plane remained aloft. Then β_i is the wing width and γ_i is the wing length, both measured in millimeters.

$$
\begin{aligned}
y_i|\mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma) & (1)\\
\mu_i &= \alpha + \beta_i + \gamma_i & (2)\\
\alpha &\sim \text{Normal}(0, 2.5) & (3)\\
\beta &\sim \text{Normal}(0, 2.5) & (4)\\
\gamma &\sim \text{Normal}(0, 2.5) & (5)\\
\sigma &\sim \text{Exponential}(1) & (6)
\end{aligned}
$$

We run the model in R using the `rstanarm` package of Brilleman et al. (2018). We use the default priors from `rstanarm`.

3.1.1 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. In particular…

We can use maths by including latex between dollar signs, for instance θ.

4 Results

Our results are summarized in **?@tbl-modelresults**.

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

B.2 Diagnostics

References

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

Brilleman, SL, MJ Crowther, M Moreno-Betancur, J Buros Novik, and R Wolfe. 2018. "Joint Longitudinal and Time-to-Event Models via Stan." https://github.com/stan-dev/stancon_talks/.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Toronto Public Library. 2024. "Library Branch General Information." https://open.toronto.ca/dataset/library-branch-general-information/.

Wickham, Hadley. 2011. "Testthat: Get Started with Testing." *The R Journal* 3: 5–10. https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.