Understanding Subsidy Allocation in Toronto's Child Care Centres: Insights from a Logistic Regression Analysis*

Non-Profit Governance, Program Participation, and Capacity Drive Funding Decisions

Claire Ma

December 10, 2024

This study examines the factors influencing government subsidy allocation to licensed child care centres in Toronto based on Licened child Care Centers from Open Data Toronto. By employing a logistic regression model, we found that non-profit governance, participation in the Canada-Wide Early Learning and Child Care (CWELCC) program, and larger capacity significantly increase a centre's likelihood of receiving subsidies. These findings highlight how funding priorities align with public goals to expand access to high-quality child care, especially for vulnerable populations. By uncovering patterns in subsidy distribution, this research provides critical insights for policymakers to ensure more equitable and effective resource allocation.

1 Introduction

Child care subsidies play a critical role in making high-quality early childhood education and care accessible to families and communities. In Toronto, licensed child care centres serve as essential providers, offering regulated and professional environments that support child development. Subsidies for licensed child care centres help close the affordability gap by offsetting the high costs of quality care, making it accessible to more families (Cleveland and Krashinsky 2009). These subsidies not only ease the financial burden on families but also ensure that children have access to nurturing environments that foster cognitive, social, and emotional growth during their formative years (Vines 2020). Understanding the factors that determine

*Code and data are available at: [https://github.com/ClaireMa0311/Toronto-licensed-child-centers.git].

which licensed centres receive subsidies is crucial for policymakers and stakeholders to promote equitable resource allocation and maximize the benefits of early childhood programs.

Licensed child care centres in urban settings like Toronto play a pivotal role in delivering high-quality, structured child care. These centres adhere to stringent regulations, ensuring compliance with standards for safety, staffing, and curriculum. Research highlights that children attending licensed centres, particularly those supported by subsidies, experience better developmental outcomes. Subsidized centres provide professional environments with trained educators, comprehensive programming, and age-appropriate resources, offering children a strong foundation for lifelong learning and success. Subsidies are a cornerstone of this ecosystem, enabling licensed centres to cover operational costs, retain qualified staff, and maintain compliance with regulatory standards, all of which enhance the quality of care provided.

Despite their importance, disparities in the allocation of subsidies remain a significant concern. Research indicates that centres in certain neighborhoods or serving specific populations may receive fewer subsidies, even when demand is high (Johnson, Ryan, and Brooks-Gunn 2012). Additionally, characteristics such as enrollment capacity, accreditation status, and program focus (e.g., infant care versus pre-kindergarten) often influence a centre's eligibility and prioritization for funding. These discrepancies highlight the need for a data-driven approach to understanding and improving the distribution of subsidies among licensed child care centres, ensuring that resources are allocated equitably to maximize their impact.

This study utilizes the Toronto Open Data: Licensed Child Care Centres dataset to explore the factors that affect subsidy allocation to licensed centres. By analyzing variables such as ward, operating auspice (Commercial, Non Profit or Public), CWELCC, type of building, and total space, this research aims to uncover the question - "What factors influence the allocation of subsidies to licensed child care centres in Toronto?". The findings will contribute to informing policies that promote equity and efficiency in subsidy allocation, ultimately supporting the goals of accessible and high-quality child care in Toronto. Additionally, these insights can serve as a valuable tool for child care centres to self-assess and enhance their eligibility for subsidies.

Using a logistic regression model, the analysis identifies three key predictors of subsidy allocation: non-profit governance, participation in the Canada-Wide Early Learning and Child Care (CWELCC) program, and licensed capacity. The findings reveal that centres operated by non-profit organizations are significantly more likely to receive subsidies, reflecting a policy preference for supporting entities that prioritize social objectives over profit motives. Participation in the CWELCC program emerges as the strongest predictor, highlighting the importance of alignment with affordability initiatives in determining funding priorities. Additionally, larger centres with greater licensed capacity are moderately more likely to receive subsidies, indicating that operational scale plays a role in funding decisions, though it is less significant than governance and policy alignment. These findings underscore the need for policies that address disparities in subsidy distribution while ensuring that funding supports accessible, high-quality child care for families across Toronto.

The remainder of this paper is structured as follows. Section 2 describes the dataset used for the analysis, including its source, key variables, and preprocessing steps, emphasizing its real-world context and relevance to understanding subsidy allocation in Toronto's child care system. Section 3 details the logistic regression model employed in the study, including the rationale for its selection, mathematical formulation, and key predictors. The model assumptions, diagnostics, and validation methods are also discussed to ensure robustness and reliability. Section 4 contains the results of the logistic regression analysis are presented, highlighting the significant predictors of subsidy allocation, ass well as the model's fit. Section 5 discusses what was done in this study, giving detailed explanation of implications of the findings, limitations and suggestion for future studies. Furthermore, an appendix including supplementary materials and data collection evaluation is provided to ensure transparency and reproducibility.

2 Data

2.1 Data Overview

The dataset used in this study was sourced from the (Gelfand 2022; **toronto2024licensed?**) made publicly available by the City of Toronto. This original raw dataset provides detailed information about licensed child care centres, including 20 variables capturing aspects of their location, operating auspice (e.g., non-profit, public, or commercial), space usage, building type, participation in government programs such as the Canada-Wide Early Learning and Child Care (CWELCC) system, and other operational details. These data offer valuable insights into the factors influencing subsidy allocation, a key policy tool for improving access to early childhood education and care. By translating real-world phenomena into structured data entries, this dataset enables a comprehensive exploration of equity and efficiency in child care funding. Detailed data collection analysis is in the Appendix. Table 1 visualized first 10 rows of cleaned data to be used in this study.

Figure 1 visualizes the distribution of subsidies among licensed child care centres in Toronto. The x-axis represents the subsidy status (0 = No, 1 = Yes), while the y-axis shows the count of centres in each category. This distribution highlights the government's prioritization of subsidies to a majority of licensed centres, reflecting efforts to enhance accessibility and affordability. However, the remaining non-subsidized centres indicate potential gaps or disparities in resource allocation, which could be further explored to ensure equitable funding.

2.2 Method

The dataset used for this study is the `Licensed Child Care Centres dataset`, sourced from the (Gelfand 2022; **toronto2024licensed?**) It provides detailed information about licensed child care centres in Toronto, capturing aspects such as their governance, capacity, infrastructure, and subsidy status. The original dataset consisted of 1071 records for all licensed

Table 1

Overview of Cleaned Data

Displaying the first 10 rows

ward	AUSPICE	bldg_type	cwelcc_flag	TOTSPACE	subsidy
3	Non Profit Agency	Public Elementary School	1	164	1
8	Non Profit Agency	Public Elementary School	1	83	1
25	Non Profit Agency	Catholic Elementary School	1	102	1
10	Non Profit Agency	Other	1	65	1
20	Non Profit Agency	High Rise Apartment	1	26	1
24	Non Profit Agency	Community College/University	1	62	1
6	Non Profit Agency	Public High School	1	49	1
24	Commercial Agency	High Rise Apartment	1	46	1
19	Non Profit Agency	Public Elementary School	1	51	1
8	Non Profit Agency	Public Elementary School	1	153	1

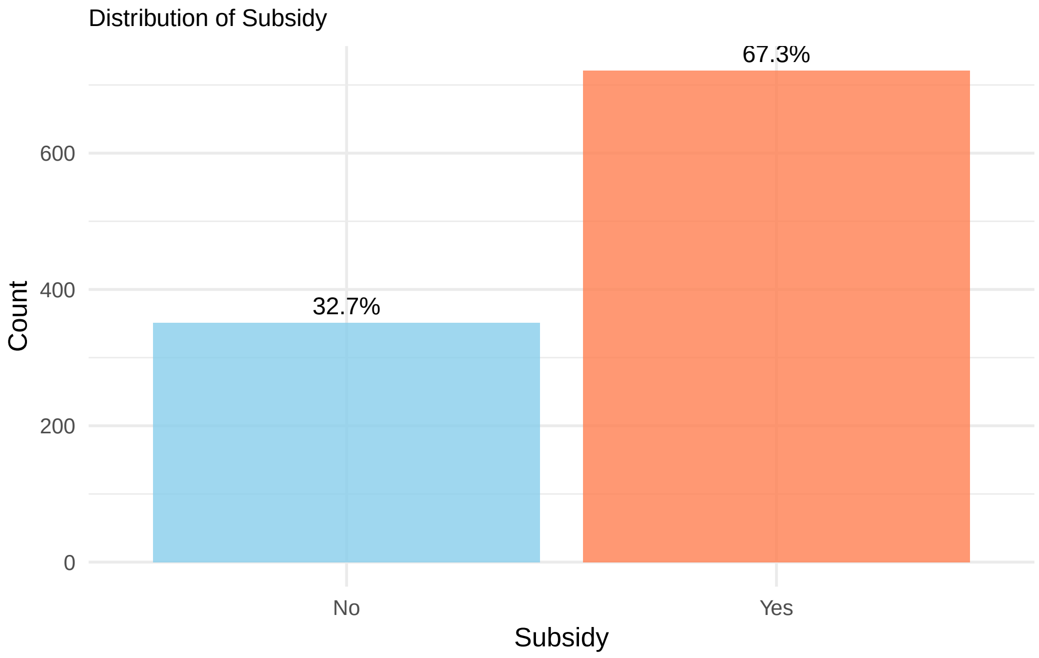

Figure 1: Distribution of Subsidy Allocation: Disparities Exist

child care centres within the city. For this analysis, the data underwent preprocessing to focus on variables relevant to the study, such as subsidy status, building type, CWELCC participation, total space, and operating auspice. These variables were retained to examine how different factors influence subsidy allocation. The dependent variable **`subsidy`**, originally recorded as "Yes"/"No," was encoded as 1 for subsidized and 0 for non-subsidized centres. Similarly, the **CWELCC participation** variable was converted into a binary format (1 =Y, 0 =N). Categorical variables such as **building type** and **`operational AUSPICE`** were consolidated to simplify analysis and address sparse categories. Additionally, missing and irrelevant records were removed to ensure the dataset was both accurate and meaningful for the research objectives.

The data for this study was systematically downloaded, cleaned, analyzed, and visualized using `R` (R Core Team 2023), a statistical programming language. The following are major packages used for this study:

- `opendatatoronto`(Gelfand 2022): Used to access and retrieve the Licensed Child Care Centres dataset directly from the City of Toronto's open data portal.

-`arrow` (Richardson et al. 2024): Provided efficient tools for reading and writing Parquet files, enabling fast and memory-efficient handling of large datasets during analysis.

- `here` (**citehere?**): Simplified file referencing by creating relative paths, ensuring reproducibility and consistency in accessing datasets, scripts, and outputs across different working environments.
- `readr` (Wickham, Hester, and Bryan 2024): Simplified the import and parsing of raw data into R.
- `tidyverse` (Wickham et al. 2019): Streamlined data manipulation, cleaning, and visualization processes.
- `dplyr` (Wickham et al. 2023): Provided tools for filtering, transforming, and summarizing the dataset effectively.
- `tidyr` (**citetidyr?**): Enabled data tidying processes such as reshaping, separating, or combining, and handling missing values to structure the dataset into a clean and analyzable format.
- `ggplot2` (Wickham 2016): Created powerful and flexible visualizations tailored to the analysis needs.
- `ggcorrplot` (**citeggcorrplot?**): Simplified the visualization of correlation matrices with customizable heatmaps, including options for labeling and styling to enhance interpretability and presentation.
- `glmnet` (**citeglmnet?**): Applied for fitting regularized regression models and feature selection.

- `modelsummary` (**citemodelsummary?**): Streamlined the creation of comprehensive and customizable summary tables for regression models, including key statistics such as AIC, BIC, and coefficient estimates, to facilitate clear model comparisons and reporting.
- `stats` (**citestats?**): Provided core statistical functions to fit models, perform hypothesis testing, and calculate measures such as p-values, log-likelihoods, and AIC for model evaluation.
- `car` (Fox and Weisberg 2019): Used for diagnostic tools, including Variance Inflation Factor (VIF) tests, to assess multicollinearity.
- `caret` (Kuhn and Max 2008): Enabled the development, validation, and evaluation of machine learning models, including training-test splits and performance metrics.
- `pROC` (**citeproc?**): Facilitated the computation and visualization of Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) metrics, enabling robust evaluation of the model's classification performance.
- `stargazer` (Hlavac 2022): Generated formatted regression tables for outputs.
- `broom` (**citebroom?**): Simplified the process of converting model outputs into tidy data frames, enabling seamless integration of statistical results with visualization and reporting workflows.
- `kableExtra` (**citekableExtra?**):Enhanced the presentation of tables by adding advanced formatting options, including captions, headers, and styling, for polished and professional reporting.
- `gt` (**citegt?**): Enabled the creation of aesthetically pleasing and highly customizable tables for presenting data and model summaries in a clean and professional format.
- `knitr` (Xie 2021): Dynamically integrated code, results, and plots into the final document for seamless reporting.

2.3 Measurement

This analysis focuses on the following variables, with a specific emphasis on subsidy as the dependent variable:

- `Subsidy`: The binary dependent variable indicating whether a licensed child care centre receives a government subsidy.
- 1: The centre is subsidized.
- 0: The centre is not subsidized.
- `ward`: 'A numeric variable representing the ward number for child care centres.
- `Operating Auspice`: The operating auspice of the child care centre, describing its governance and operational model. Possible values include:

- Non-Profit: Centres operated by non-profit organizations, often reinvesting surplus revenues into quality improvements.
- Commercial: For-profit centres operated by private organizations.
- Public: Centres run by public agencies or school boards.
- `Building Type`: The type of building where the child care centre operates, reflecting its infrastructure.

Examples include:

- Commercial Building
- Community College/University
- Community Health Centre
- Community Rec/Centre - Board Run
- Community/Rec Centre - City
- Community/Recreation Centre
- High Rise Apartment
- Hospital/Health Centre
- House
- Industrial Building
- Low Rise Apartment
- Office Building
- Place of Worship
- Private Elementary School
- Public (school closed)
- Public Elementary Special
- Public High School
- Public Middle School
- Purpose Built
- Synagogue
- Other
- `CWELCC Participation`: A binary variable indicating participation in the Canada-Wide Early Learning and Child Care (CWELCC) program:
- 1: The centre participates in CWELCC, enabling reduced child care fees.
- 0: The centre does not participate in CWELCC.
- `Total Space`: A numerical variable representing the total licensed capacity (spaces available) for all age groups at a child care centre.

Detailed information about these variables' information and data structure is presented in Table 1.

The variables were carefully selected based on literature-supported relevance to subsidy allocation and their representation of real-world phenomena

Figure 2 illustrates the distribution of subsidy status (1 = Subsidized, 0 = Not Subsidized) across various building types housing licensed child care centres. Notably, Public Schools, Purpose-Built Facilities, and Community Recreation Centres exhibit higher proportions of subsidized centres. These facilities are often designed to meet regulatory requirements for child care, including adequate space, safety standards, and accessibility, aligning closely with subsidy allocation policies (Cleveland and Krashinsky 2009). Conversely, building types such as Industrial Buildings, Private Elementary Schools, and Office Buildings show lower proportions of subsidized centres, likely due to infrastructure challenges or misalignment with subsidy eligibility criteria, such as limited accessibility or higher operational costs (**yan2011impact?**). These patterns suggest that building type significantly influences subsidy distribution. Given the numerous categories of building types, a de tailed analysis is warranted to fully understand these trends.

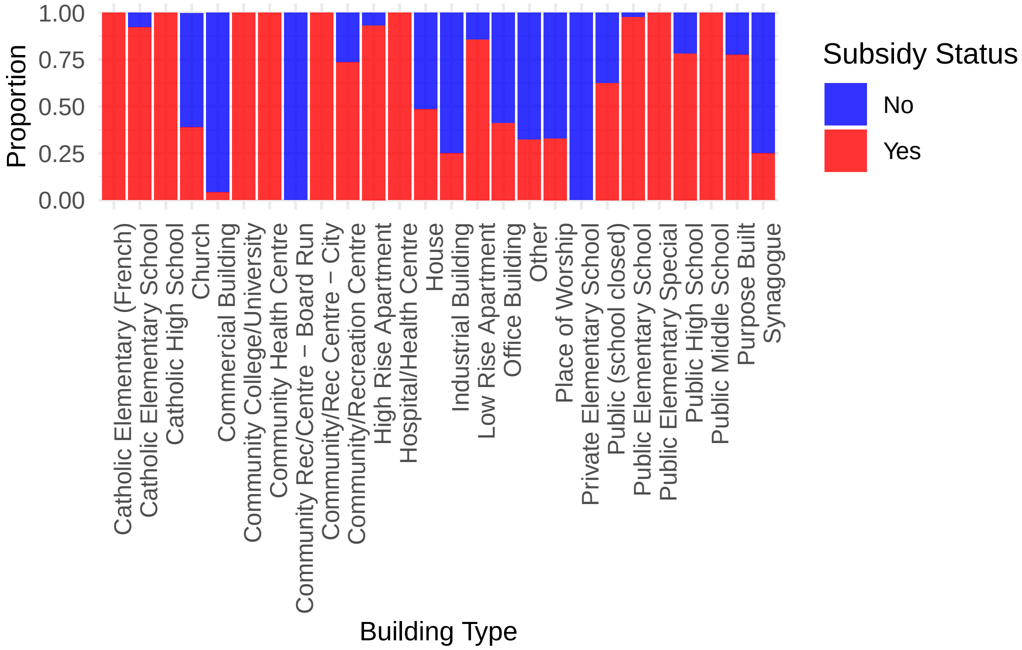

Figure 2: Subsidy Allocation by Building Type: Higher Subsidy Proportions in Public and Purpose-Built Facilities

Figure 3 illustrates the proportional distribution of subsidy status (1 = Subsidized, 0 = Not Subsidized) across the different operating auspices of licensed child care centres: Commercial Agency, Non-Profit Agency, and Other. Non-Profit Agencies and "Other" entities are predominantly subsidized, while Commercial Agencies display a more balanced distribution. This aligns with research indicating that non-profits rely heavily on subsidies to deliver public goods and services, as they often operate in markets with limited profitability (H. B. Hansmann 1979). Conversely, commercial entities are less reliant on subsidies due to their revenue-driven models.The dominance of subsidies in the "Other" category suggests this group may include hybrid or public-private organizations aligned with specific government initiatives (Anheier 2014). Such reliance reflects the strategic use of subsidies to support services underserved by the private market (Burton A. Weisbrod 2000).

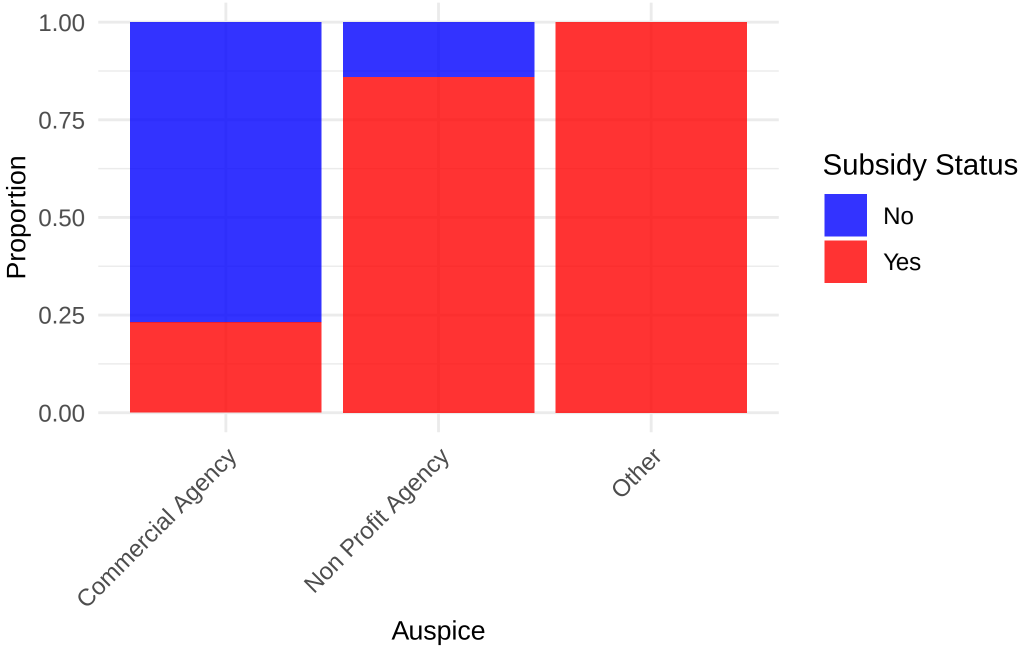

Figure 3: Subsidy Allocation by Operating Auspice: Non-Profit Agencies and 'Other' Predominantly Subsidized

Figure 4 a heatmap illustrates the correlations between three variables: total space, CWELCC participation and subsidy. The positive correlation between total space and subsidy (0.25) suggests that larger facilities are modestly more likely to receive subsidies, reflecting their capacity to serve larger populations or provide greater public benefits. This aligns with research indicating that larger organizations often have the resources and visibility to secure subsidies (H. Hansmann 1980; **salamon2002tools?**). Additionally, the stronger correlation between CWELCC participation and subsidy (0.48) highlights that subsidy allocation may target entities meeting specific programmatic or policy criteria, consistent with the literature emphasizing strategic targeting of subsidies to maximize societal impact (Burton A. Weisbrod 1998). The weaker correlation between total space and CWELCC oarticipation (0.16) suggests that program eligibility is less dependent on size and more on qualitative factors like service type or demographic focus, which is supported by (**anheier2005nonprofit?**) analysis of nonprofit funding models. Together, these correlations emphasize the nuanced role of subsidies in balancing operational scale and policy alignment, underscoring the importance of strategic allocation in public funding (Salamon 1995).

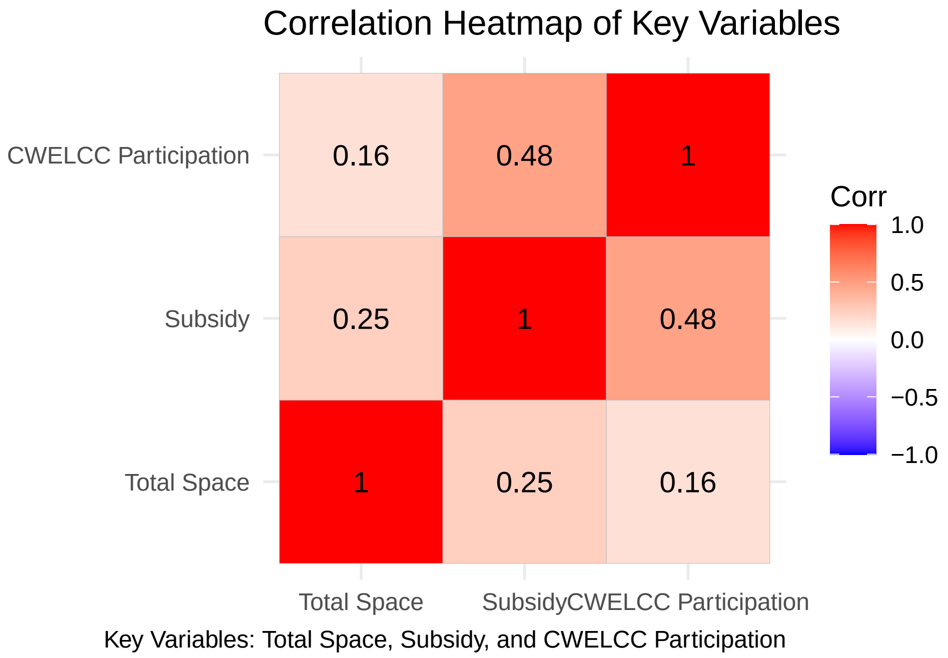

Figure 4: Correlation Between Total Space, Subsidy Allocation, and CWELCC Participation

3 Model

3.1 Model Specification

To investigate the factors influencing subsidy allocation to licensed child care centres, a reduced logistic regression model was specified. The dependent variable,`Subsidy`, is a binary indicator representing whether a child care centre receives government subsidy (1 = subsidized, 0 = not subsidized). The model includes three key predictors:

`Operating Auspice`: A categorical variable indicating the governance model of the child care centre (e.g., Non-Profit Agency, Other). Non-profit agencies are hypothesized to be positively associated with subsidy allocation, consistent with previous research emphasizing their prioritization in funding schemes (**cleveland2005benefits?**).

`CWELCC participation`: A binary variable capturing whether the centre participates in the Canada-Wide Early Learning and Child Care program (1 = participates, 0 = does not participate). Centres participating in this initiative are expected to have higher odds of receiving subsidies due to their alignment with government objectives of affordability and accessibility (**friendly2022affordable?**).

`Total Space`: A continuous variable representing the number of licensed spaces available in the centre. Larger centres are hypothesized to have higher odds of subsidy allocation, as they can accommodate more families and align with policy goals of maximizing access (H. Hansmann 1980; **salamon2002tools?**).

The logistic regression model can be expressed mathematically as follows:

$$\text{logit}(P(Y=1)) = \beta_0 + \beta_1(\text{Non-Profit Auspice}) + \beta_2(\text{Other Auspice}) + \beta_3(\text{CWELCC Participation}) + \beta_4(\text{Total Spa}$$

Where:

- $P(Y=1)$: The probability that a child care center receives a subsidy ($Y=1$).
- $\text{logit}(P(Y=1))$: The log-odds of receiving a subsidy, defined as:

  $$\text{logit}(P(Y=1)) = \ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right)$$

  This transformation ensures that the predicted probabilities remain between 0 and 1.

3.1.1 Parameters:

1. β_0 **(Intercept):**
   - Represents the baseline log-odds of subsidy allocation when all predictors are set to their reference categories or zero values.
   - It provides the starting point for predicting subsidy probabilities.
2. β_1 **(Non-Profit Auspice):**
   - Measures the effect of a center being operated as a non-profit compared to the baseline (e.g., commercial auspice) on the log-odds of subsidy allocation.
   - A positive β_1 increases the likelihood of receiving a subsidy for non-profit centers relative to commercial centers.
3. β_2 **(Other Auspice):**
   - Measures the effect of a center being operated as "other type" compared to the baseline (e.g., commercial auspice) on the log-odds of subsidy allocation.
   - A positive β_2 increases the likelihood of receiving a subsidy for other-type centers relative to commercial centers.
4. β_3 **(CWELCC Participation):**
   - Captures the influence of participating in the Canada-Wide Early Learning and Child Care (CWELCC) program on the log-odds of subsidy allocation.

5. β_4 **(Total Space):**

    - Represents the effect of the number of licensed child care spaces (a continuous variable) on the log-odds of subsidy allocation.

3.2 Model Justification

The logistic regression model was chosen for this analysis because it is well-suited to the binary nature of the dependent variable, subsidy status (1 = Subsidized, 0 = Not Subsidized). A logistic regression model using the binomial family is specifically designed to model dichotomous outcomes by estimating the log-odds of the event occurring as a linear function of predictor variables (Hosmer, Lemeshow, and Sturdivant 2013). The binomial family is appropriate here because it assumes that the dependent variable follows a Bernoulli distribution, where each observation represents a binary outcome (subsidized or not subsidized). This ensures that the predicted probabilities remain between 0 and 1, aligning with the real-world constraints of the problem. Additionally, logistic regression provides interpretable coefficients, which indicate the direction and magnitude of the relationship between each predictor and the log-odds of subsidy allocation. This makes it particularly useful for guiding policy decisions, as coefficients can be directly converted into odds ratios for actionable insights (Peng, Lee, and Ingersoll 2002).

The modeling process follows a detailed and methodical approach to ensure the development of a robust logistic regression model for predicting subsidy allocation in licensed child care centers. Initially, the categorical variables operating auspice and building type were encoded as factors to ensure proper interpretation of their qualitative nature by the model. The dataset was then split into training (70%) and testing (30%) sets, allowing for model validation on unseen data to assess generalizability. A full logistic regression model was first fit to include all predictors (subsidy ~ .), providing a baseline for further analysis. To refine the model, stepwise selection based on the Akaike Information Criterion (AIC) was employed, balancing goodness of fit with model parsimony by penalizing unnecessary complexity. This process resulted in the selection of the most significant predictors for subsidy allocation.

To further simplify and enhance interpretability, the ward variable was excluded because it lacked statistical significance and added redundancy, as its effects were captured by other predictors like CWELCC participation and total space. Additionally, ward had no strong theoretical justification as a direct determinant of subsidy allocation. The building type variable was removed due to high dimensionality, sparse representation in many categories, and statistical insignificance. Its effects are likely mediated by other variables, such as total space and auspice. Excluding it improved parsimony and interpretability.

The final model retained operating auspice, CWELCC participation, and total space, as these predictors are strongly supported by theory and data. This approach balances simplicity and accuracy, ensuring the model remains relevant for informing equitable subsidy allocation policies.

Multiple models, including the full model, the AIC-based model, and the simplified model, were compared using goodness-of-fit metrics such as AIC and Deviance to ensure the best-performing model was selected. Additionally, McFadden's R-squared was calculated to evaluate the explanatory power of the model, comparing the log-likelihood of the fitted model to a null model. A higher McFadden's R-squared value confirmed the model's ability to explain a significant portion of the variance in subsidy allocation.

By training and testing the models on separate subsets of the data, the process ensured validation and minimized the risk of overfitting. The final model was chosen for its simplicity, interpretability, and ability to retain statistically significant predictors that align with theoretical expectations. This rigorous process not only ensures robust prediction but also aligns with real-world phenomena in subsidy allocation, making the model both practical and reliable.

3.3 Model Assumptions

Detailed model assumption check is in Section A.

3.4 Alternative Models

Alternative models like decision trees and random forests were considered but were found less interpretable for policy-focused analyses. Logistic regression was chosen for its balance of simplicity, interpretability, and effectiveness.

4 Results

The logistic regression model's performance metrics are presented in Table 2.

The regression model reveals several key observations regarding the predictors and overall goodness of fit. The intercept, estimated at -4.818, represents the baseline log-odds of the outcome when all predictors are at their reference or zero level. While not directly interpretable, it serves as a baseline reference for the model. Among the predictors, the operating auspice: non-profit Agency category significantly increases the log-odds of the outcome, with an estimate of 2.937 and a small standard error of 0.192, indicating a strong and reliable positive effect. However, the operating auspice: other category shows an unusually large coefficient (18.680) paired with a very high standard error (631.207), suggesting instability. The CWELCC participation variable also exhibits a strong and reliable positive effect, with an estimate of 3.101 and a standard error of 0.346, indicating that being flagged as CWELCC significantly increases the log-odds of the outcome. Additionally, the total space variable has a small but statistically significant effect, with an estimate of 0.013 and a low standard error of 0.003, reflecting robustness. In terms of model fit, the dataset includes 750 observations, and the model's AIC (760.8) and BIC (785.7) suggest a reasonable balance between goodness of fit and complexity, as lower

Table 2: Regression Result

	Final Model
(Intercept)	−4.818
	(0.399)
	(<0.001)
AUSPICENon Profit Agency	2.937
	(0.192)
	(<0.001)
AUSPICEOther	18.680
	(631.207)
	(0.976)
cwelcc_flag	3.101
	(0.346)
	(<0.001)
TOTSPACE	0.013
	(0.003)
	(<0.001)
Num.Obs.	1072
AIC	760.8
BIC	785.7
Log.Lik.	−375.414
RMSE	0.32

values are generally preferred (Burnham and Anderson, n.d.). The log-likelihood of -375.414 also supports an adequate fit, with a higher (less negative) value indicating better alignment between the model and the data. Finally, the RMSE of 0.32 reflects the model's predictive accuracy, with a low value indicating that the model's predictions closely match the observed data. Overall, the model performs well but may require refinement, particularly in addressing instability in the AUSPICE: Other variable. Figure 5 further visualizes the significance of predictors.

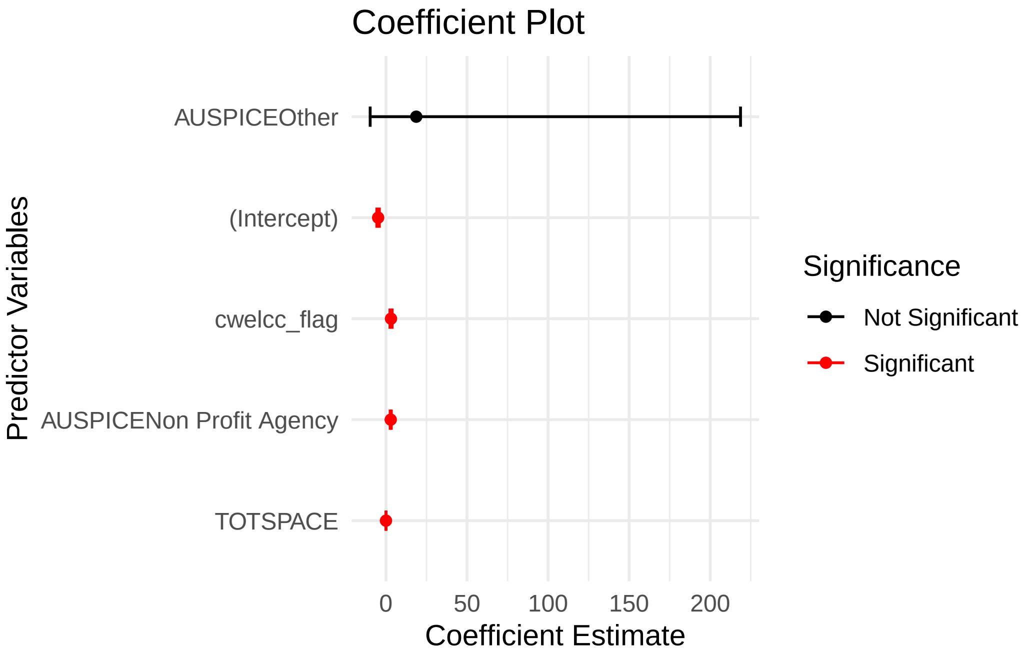

Figure 5: Coefficient Estimates for Predictors of Subsidy Allocation

The evaluation of the logistic regression model shows strong performance based on McFadden's R^2 and the Area Under the Receiver Operating Characteristic Curve (AUC). McFadden's R^2 which is a pseudo-R^2 metric specifically designed for logistic regression, was calculated as 0.454. This value suggests that the model explains 45.4% of the variance in the outcome variable, indicating a well-fitting model. Values between 0.2 and 0.4 are considered indicative of a good model fit, and values above 0.4 demonstrate an excellent fit. Thus, the model's R^2 value strongly supports its utility in predicting the outcome.

The ROC curve in Figure 6 shown further validates the model's classification ability, with an AUC of 0.906. An AUC value close to 1 indicates excellent discriminative power, where the model effectively separates true positives from false positives. The AUC of 0.8998 places this model on the borderline of very good and excellent discrimination, underscoring its robust predictive capacity.

Figure 7 provides a visual representation of the alignment between the predicted probabilities generated by the logistic regression model and the actual observed binary outcomes. This

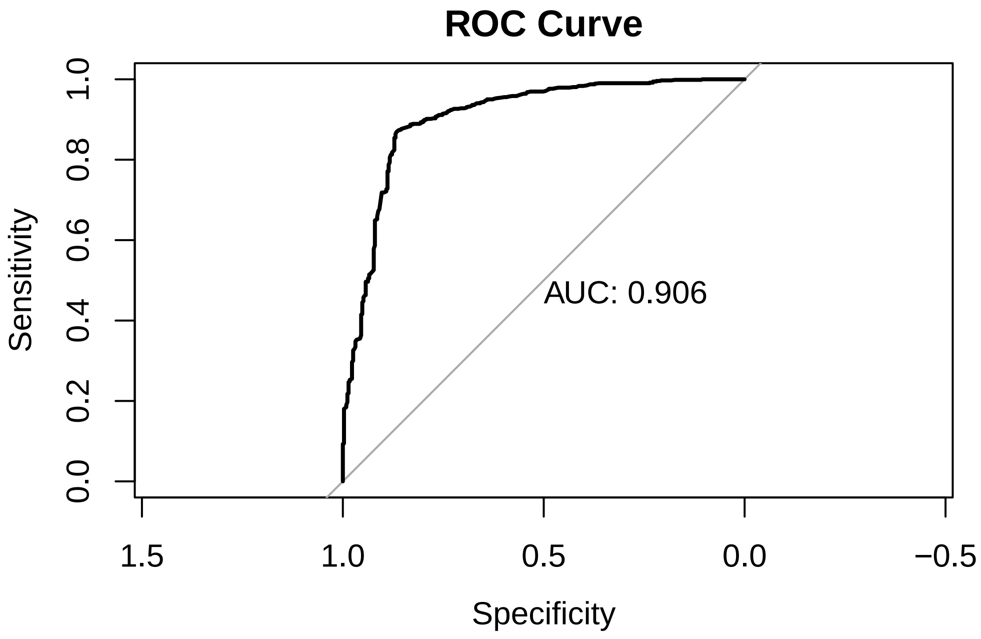

Figure 6: ROC Curve for Model Performance: Strong Model Performance with AUC 0.906

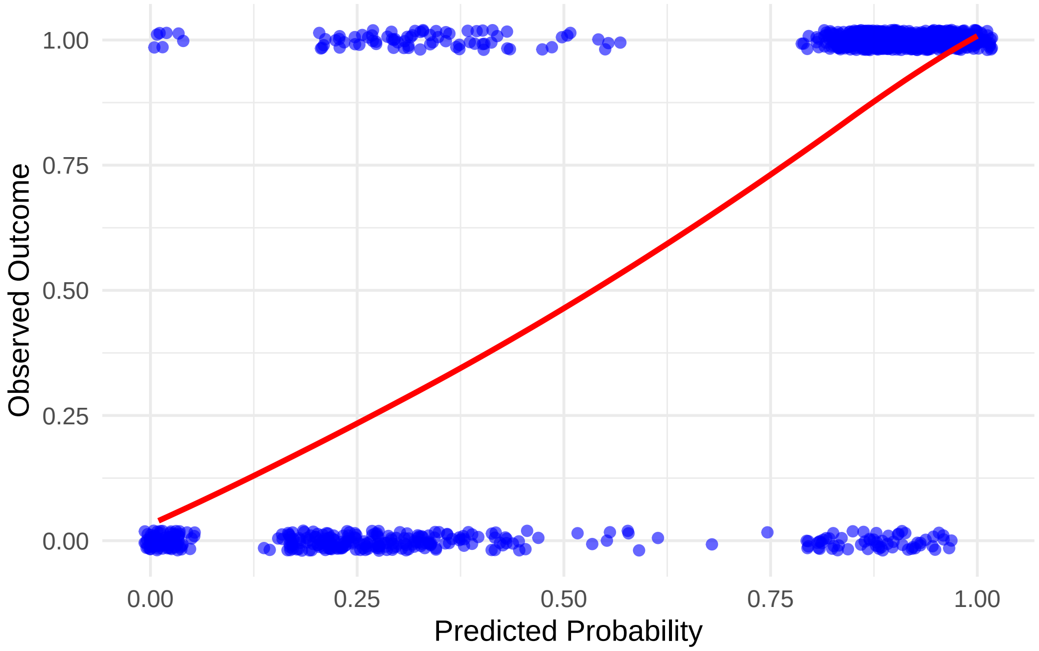

Figure 7: Predicted vs. Observed Probability Plot: Model Calibration and Accuracy

diagnostic tool is crucial for assessing the model's calibration and predictive accuracy, as emphasized by Harrell (Frank E. Harrell 2012). The x-axis shows the predicted probabilities ranging from 0 to 1, while the y-axis represents the observed outcomes, where 0 indicates the absence of an event and 1 indicates its presence. The red diagonal line serves as the reference for perfect calibration, where predicted probabilities perfectly match observed outcomes.

The clustering of blue points near 0 and 1 along the y-axis indicates that the model effectively distinguishes between the two classes, a key goal in binary classification modeling. This observation aligns with the high AUC value of 0.8998 observed in the ROC curve, which demonstrates the model's excellent discriminative ability. Additionally, points near the diagonal line further suggest strong calibration, where the predicted probabilities closely align with actual outcomes. The concentration of blue points at the extremes of 0 and 1 highlights that the model makes confident and accurate predictions in many cases

However, some dispersion is observed at lower predicted probabilities, particularly around 0.25 and 0.5. These deviations may reflect cases where the model struggles to classify outcomes or makes less confident predictions. Such discrepancies could indicate areas where further refinement of the model or predictors is necessary. These deviations could also be linked to specific subgroups or predictor interactions, necessitating additional diagnostics or recalibration (Steyerberg et al. 2010).

Overall, this graph demonstrates that the model is well-calibrated and effective in its predictions, with minor limitations at lower probability levels. The close alignment of predictions with observed outcomes supports the model's reliability in classification tasks. This interpretation aligns with other metrics, such as McFadden's R^2 and AUC, which indicate robust overall model performance while pointing to opportunities for refinement. Including this visualization in the results section provides a clear and data-driven assessment of the model's strengths and areas for potential improvement.

5 Discussion

Subsidy allocation to licensed child care centres in Toronto addresses the critical issue of affordability, making high-quality early childhood education accessible to more families as licensed centres must meet strict safety, staffing, and programming standards, which foster children's cognitive, social, and emotional growth (Cleveland and Krashinsky 2009). Without subsidies, the cost of regulated care can be prohibitive, particularly for low-income households. This paper undertakes a detailed examination of the factors influencing subsidy allocation to licensed child care centres in Toronto, leveraging the City of Toronto's Licensed Child Care Centres dataset from Open Data porta, which includes detailed information about 1,071 licensed centers(Gelfand 2022). This study employs a logistic regression model-binomial family to identify relationships between key predictor variables and the binary dependent variable – receiving subsidies.

The dataset underwent meticulous preprocessing to ensure reliability and accuracy for analysis. This involved several key steps of data cleaning, variable transformation and training-test Split. Initially, variables of operating auspice, ward, building type, total space and CWELCC participation were chosen based their relevance and practical influence subsidy outcomes and forecasting supported by scholar literatures. After fitting and improving the model using stepwise selection based on the Akaike Information Criterion (AIC), non-significant variables (ward and building type) were removed. The final model retained operating auspice, CWELCC participation, and total space as key predictors. The assumptions underpinning the logistic regression model were thoroughly validated, ensuring its statistical rigor and reliability. Independence of observations was confirmed, as each data point corresponds to a unique child care centre, eliminating concerns of clustering or dependency. Linearity between continuous predictors, such as total space, and the logit transformation was validated through component + residual plots, which revealed consistent patterns, supporting the appropriateness of the model. Additionally, multicollinearity was minimized, with Variance Inflation Factor (VIF) values well below the critical threshold of 5, indicating that the predictors are independent and the coefficient estimates are stable. Together, these validations affirm the robustness of the model and its suitability for the analysis.

The result coefficients in table the logistic regression model provide a detailed understanding of how key predictors influence the likelihood of subsidy allocation to licensed child care centres in Toronto. By measuring the change in log-odds of subsidy allocation associated with each predictor, the model quantifies the relative importance of governance models, program participation, and operational capacity. The coefficient for non-profit auspice emerged as strongly positive and statistically significant, indicating that non-profit child care centres are far more likely to receive subsidies than commercial or public centres. This is reflective of policy priorities that favor organizations with a social mission over those with profit-oriented objectives. Non-profits are often viewed as better aligned with public goals of equity and accessibility, as they reinvest surplus revenues into improving care quality rather than distributing profits (H. Hansmann 1980). This is consistent with research by (Salamon 1995), who noted that non-profits often fill critical gaps in the provision of public services, making them natural candidates for targeted funding. In the context of child care, non-profits may also be more likely to serve economically vulnerable populations, further justifying their prioritization for subsidies. The odds ratio derived from this coefficient underscores the strength of this relationship, suggesting that non-profit centres are several times more likely to receive subsidies compared to commercial centres. Participation in the Canada-Wide Early Learning and Child Care (CWELCC) program was the most significant predictor in the model, with a highly positive coefficient. This result highlights the critical role of policy alignment in subsidy allocation. Centres participating in CWELCC demonstrate compliance with national and provincial affordability initiatives, making them high-priority recipients of government funding. Research by (Bennett and Moss 2011) supports this finding, emphasizing that early childhood programs aligned with broader governmental objectives are often better positioned to secure resources. The odds ratio for CWELCC participation indicates that such centres are substantially more likely to receive subsidies, which is consistent with targeted funding strategies aimed at expanding access to high-quality, affordable child care (Kaga, Bennett, and Moss 2010). Total space, representing the number of spaces a centre is licensed to operate, was another important predictor, although its coefficient was smaller in magnitude compared to operating auspice and CWELCC participation. This positive coefficient suggests that larger centres are more likely to receive subsidies, likely due to their capacity to serve more families and their operational scale. Larger centres can achieve economies of scale and often have greater visibility, making them attractive candidates for public funding (Burton A. Weisbrod 1998). However, the smaller effect size of this variable indicates that capacity alone is insufficient to determine subsidy allocation. This finding aligns with (Penn 2011), who argues that while capacity is an important practical consideration, other factors—such as governance and programmatic alignment—carry more weight in funding decisions. Interpreting these coefficients through odds ratios further clarifies their impact. For instance, the odds ratio for CWELCC participation suggests that centres in the program are several times more likely to receive subsidies compared to non-participating centres, holding other factors constant. Similarly, the odds ratio for non-profit governance reinforces the preferential treatment of non-profits in subsidy allocation policies. These findings align with the broader literature on public funding, which emphasizes the strategic use of subsidies to support centres that meet both operational and policy criteria (Penn 2011). The statistical significance of these coefficients underscores their reliability as shown in figure. Variables such as non-profit auspice and CWELCC Participation demonstrated small standard errors and strong significance levels, reinforcing their influence on subsidy allocation. In contrast, variables with less consistent effects, such as certain building types or governance categories in the "other" category, exhibited instability, suggesting that their role in funding decisions may be more context-specific or secondary. The model's findings highlight a strategic approach to subsidy allocation, where funding decisions are guided by alignment with policy goals, operational capacity, and governance structures.

The logistic regression model applied in this research proves to be an ideal tool for analyzing subsidy allocation to licensed child care centres in Toronto, particularly due to its capacity to handle binary outcomes effectively. By modeling the log-odds of an event, logistic regression ensures predicted probabilities remain between 0 and 1, aligning seamlessly with the binary nature of the dependent variable—whether a centre receives a subsidy. This feature makes the model both statistically appropriate and easy to interpret. A notable strength of the model lies in its interpretability. Coefficients can be transformed into odds ratios, providing a clear understanding of how predictors influence the likelihood of receiving subsidies. For example, the positive coefficients for non-profit auspice and CWELCC Participation underscore their significant, policy-relevant impact on subsidy distribution. This clarity makes the model particularly valuable for stakeholders and policymakers who require actionable and easily communicated insights. The model's performance metrics further demonstrate its effectiveness. A McFadden's R^2 value of 0.454 indicates that the model explains a substantial portion of the variance in subsidy allocation. Additionally, the Area Under the Curve (AUC) of 0.906 highlights the model's exceptional discriminative ability, confirming its capacity to accurately differentiate between subsidized and non-subsidized centres. The low Root Mean Squared Error (RMSE) further attests to the model's strong predictive performance. Together, these metrics validate the model's reliability and robustness. Another key advantage of logistic regression is its simplicity compared to more complex approaches such as random forests or decision trees. While those alternatives might offer marginally higher predictive power, they lack the transparency and interpretability required for policy-focused research. Logistic regression strikes a critical balance between accuracy and clarity, making it a practical choice for guiding equitable subsidy allocation strategies. Moreover, the flexibility of logistic regression allows for extensions to explore more intricate relationships. For instance, incorporating interaction terms or temporal data could provide deeper insights while maintaining the model's usability. This adaptability ensures the model remains relevant and applicable as policy environments evolve.

Our model, while effective, has several limitations that warrant consideration. One of the primary limitations is its reliance on the assumption of linearity between predictors and the log-odds of the outcome. Although this assumption is often reasonable, it may oversimplify complex, non-linear relationships that can exist in real-world phenomena such as subsidy allocation. As (Frank E. Harrell 2015) notes, failing to capture these non-linear relationships can lead to biased estimates and limit the model's predictive power. Another limitation is the model's dependence on correct specification of predictors. Excluding relevant variables or including irrelevant ones in initial variable selection can result in omitted variable bias or overfitting, respectively. Additionally, the model does not inherently account for higher-order interactions unless explicitly included. For instance, the combined effect of Non-Profit Auspice and CWELCC Participation could provide valuable insights into how policy and governance jointly influence subsidy allocation. (Hosmer, Lemeshow, and Sturdivant 2013) highlight that logistic regression can oversimplify complex relationships if interaction terms are not considered, potentially leading to incomplete conclusions. Finally, the static nature of the model is another drawback. Logistic regression provides a snapshot of relationships at a single point in time, failing to capture temporal dynamics or policy changes that can influence subsidy allocation over time. As Singer and (Singer and Willett 2003) suggest, incorporating longitudinal data could provide a more nuanced understanding of how these relationships evolve.

Future studies could address the limitations of this research by adopting more advanced modeling techniques and incorporating additional data to better capture the complexities of subsidy allocation. Introducing non-linear models or machine learning approaches, such as random forests or gradient boosting, could uncover hidden patterns and interactions between predictors that logistic regression may overlook. Additionally, exploring interaction terms, such as the joint effects of governance models and program participation, could provide more nuanced insights into policy impacts. Longitudinal data would be particularly valuable for understanding how subsidy allocation evolves over time, especially in response to policy changes or shifts in economic conditions. Incorporating time-series or panel data analysis could reveal temporal dynamics and provide a richer understanding of causal relationships. Geographic disparities, noted but not deeply explored in this study, could also be addressed using spatial regression or multilevel models to account for regional clustering and unobserved local factors influencing subsidy distribution. Finally, integrating additional predictors, such as demographic variables, socio-economic indicators, or quality ratings of child care centres, could enhance the explanatory power of future models. These enhancements would provide a more comprehensive view of subsidy allocation and inform strategies to ensure equitable and effective distribution of resources.

Appendix

A Model Assumption

To ensure the validity of the logistic regression model, several key assumptions were assessed, including independence of observations, the appropriateness of a binary outcome, the linearity of predictors with the logit and absence of multicollinearity. The analysis integrates results from visual diagnostics, multicollinearity tests, and statistical measures.

A.0.1 1. Independence of Observations

The logistic regression model assumes that the observations are independent of each other. In this analysis, each data point corresponds to an individual child care center, ensuring independence. There is no clustering or repeated measures within the dataset, which validates this assumption.

A.0.2 2. Binary Outcome

The logistic regression model assumes a binary dependent variable. In this case, the outcome variable, `subsidy`, is binary, indicating whether a child care center receives a subsidy (1 = Subsidized, 0 = Not Subsidized). This aligns with the model's requirement, ensuring the suitability of the binomial family for fitting the data.

A.0.3 3. Linearity of Predictors with the Logit

The Figure 8 The component + residual plots evaluate the linearity of the continuous variables and the relationship between categorical predictors and the logit transformation.

- `TOTSPACE`: The relationship between the total space (TOTSPACE) and the logit appears approximately linear, as indicated by the flat, consistent pattern of residuals around the horizontal axis. This supports the assumption of linearity.
- `CWELCC Flag`: The plot for CWELCC participation shows a horizontal trend, suggesting no significant deviation from linearity. The data support the inclusion of this variable as a binary predictor.
- `AUSPICE`: The residual patterns for categorical levels of AUSPICE (e.g., "Commercial Agency" and "Other") indicate distinct clusters with consistent variability. This suggests the categorical nature of this variable does not violate linearity assumptions. Overall, the visual inspection suggests that the linearity assumption for the predictors with the logit is met.

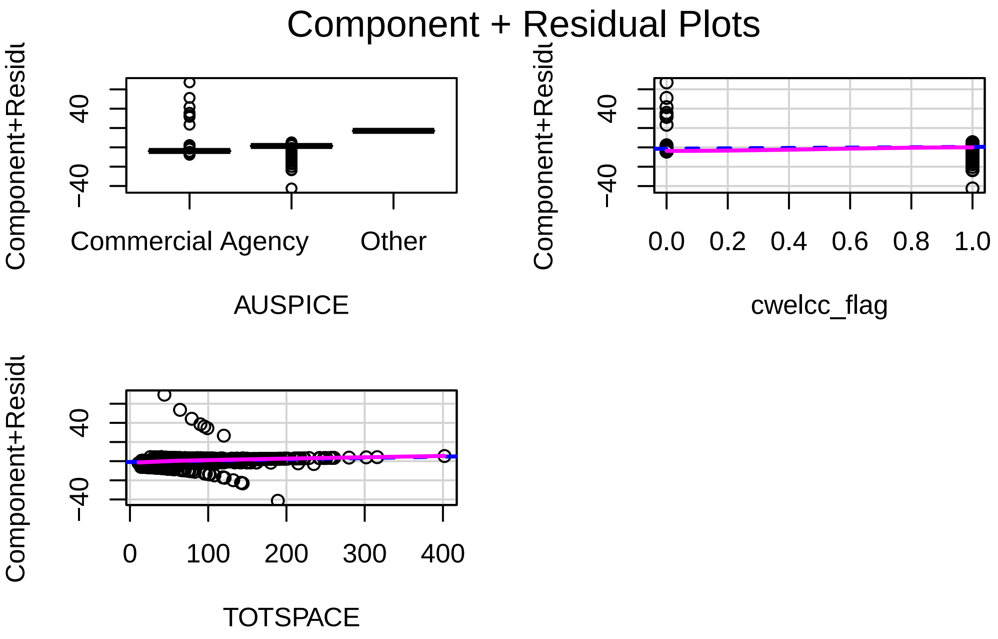

Figure 8: CR Plot for Linearity Chekck: No Violate of Linearity Assumptions

A.0.4 4. Multicollinearity Assessment

Variance Inflation Factor (VIF):

To assess multicollinearity among the predictors, the VIF values for the reduced model variables were calculated.All VIF values are below 2, well within the acceptable threshold of 5, indicating minimal multicollinearity among the predictors. The independence of the predictors ensures the stability of the coefficient estimates.

B Observational Data Methodology for Licensed Child Care Centres in Toronto

B.1 Population, Frame, and Sample

B.1.1 Population

The population includes all licensed child care centres in Toronto as recorded in the publicly available dataset provided by Open Toronto. This population aligns with studies on regulated child care, such as (Morrissey 2010), which emphasizes the role of licensing in ensuring quality and safety in early childhood education. However, the dataset excludes unlicensed child care providers, limiting its scope. Represented Groups: Centres adhering to provincial licensing requirements, as emphasized by Cleveland and Krashinsky (Cleveland and Krashinsky 2003), who note that licensed care is generally associated with higher quality standards.

Excluded Groups: Informal and unlicensed providers, which disproportionately to lower-income families and those in rural or underserved urban areas.

B.1.2 Frame

The dataset represents a comprehensive census of licensed child care centres, reflecting administrative records compiled by the city.

Strengths: Comprehensive coverage reduces selection bias and enhances geographic representativeness.

Limitations: Static data may fail to capture temporal variations or operational changes, a common limitation noted in administrative datasets.

B.1.3 Sample

The dataset uses census sampling of licensed child care centres, ensuring inclusion of all facilities within the licensing framework.

Advantages: Census data minimizes sampling error and provides a reliable basis for spatial and capacity analysis.

Drawbacks: Lack of informal care data may lead to underrepresentation of actual child care resources, echoing findings by (Tekin 2007) that informal care plays a critical role in many families' child care arrangements.

B.2 Data Collection Methodology

B.2.1 Observational Data Compilation

The dataset is derived from municipal administrative records, which are considered reliable for policy analysis (Cleveland and Krashinsky 2003). Licensing authorities collect this data as part of routine compliance monitoring, ensuring high accuracy.

Strengths: Administrative data is less prone to recall bias compared to survey data.

Weaknesses: Lack of qualitative insights into service quality or user satisfaction.

B.2.2 Non-response Handling

While the dataset does not involve survey non-response, its exclusion of unlicensed providers creates a systematic gap. Research by Fuller et al. (Fuller, Holloway, and Liang 1996) shows that informal care often fills critical voids in underserved areas, and its exclusion may understate the full scope of child care provision.

B.3 Sampling Approach and Trade-offs

B.3.1 Strengths

Comprehensive Coverage: Census sampling ensures no licensed centres are omitted. Geospatial Precision: Geographic coordinates support advanced spatial analysis, aligning with studies like (Larsen, El-Geneidy, and Yasmin 2015) on urban accessibility.

High Validity: Licensing records provide robust validity, as they are subject to regulatory verification (Cleveland and Krashinsky 2003).

B.3.2 Limitations

Exclusion of Informal Care: Research by Morrissey (Morrissey 2010) highlights that informal care is often preferred for flexibility or affordability, making its exclusion a significant limitation. Static Nature: The lack of longitudinal data limits the ability to analyze trends. Regulatory Bias: Centers excluded due to non-compliance may disproportionately represent marginalized communities.

B.4 Observational Bias and Measurement Challenges

B.4.1 Observational Bias

Biases inherent in observational datasets include selection bias (due to licensing criteria) and survivorship bias (exclusion of closed centres).

B.4.2 Measurement Challenges

Geographic Aggregation: Aggregating data to broader areas can obscure micro-level disparities. Capacity Limitations: Reported capacities may not reflect actual utilization or unmet demand, as noted by (Blau and Currie 2006).

B.5 Methodological Enhancements and Recommendations

B.5.1 Enhancements

Incorporating Temporal Data: Adding longitudinal data could enable trend analysis, echoing the approach of (Fuller, Holloway, and Liang 1996).

Integration with Demographics: Linking with census data would allow for equity analysis, consistent with the methodology used by (Larsen, El-Geneidy, and Yasmin 2015).

B.5.2 Recommendations

Addressing Informal Care: Future studies should integrate data from community surveys to include informal providers, as suggested by (Morrissey 2010).

Dynamic Updates: Regular updates to administrative records would enhance relevance, a recommendation supported.

Geospatial Enhancements: Including data on transit access or neighborhood socioeconomic conditions could enrich spatial analyses, aligning with findings by (Tekin 2007)

References

Anheier, Helmut K. 2014. *Nonprofit Organizations: Theory, Management, Policy*. Routledge.

Bennett, John, and Peter Moss. 2011. *Working for Inclusion: How Early Childhood Education and Care and Its Workforce Can Help Europe's Social and Economic Policies*. University of Edinburgh.

Blau, David M., and Janet Currie. 2006. "Pre-School, Day Care, and After-School Care: Who's Minding the Kids?" In *Handbook of the Economics of Education*, edited by Eric A. Hanushek and Finis Welch, 2:1163–1278. Elsevier. https://doi.org/10.1016/S1574-0692(06)02020-4.

Burnham, KP, and DR Anderson. n.d. "Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach2nd Ed. 2002Springer-Verlag." *New York*.

Cleveland, Gordon, and Michael Krashinsky. 2003. *The Benefits and Costs of Good Child Care*. Toronto, Canada: University of Toronto Press.

———. 2009. "The Nonprofit Advantage: Producing Quality in Thick and Thin Child Care Markets." *Journal of Policy Analysis and Management* 28 (3): 440–62.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. https://www.john-fox.ca/Companion/.

Fuller, Bruce, Susan D. Holloway, and Xiaoyan Liang. 1996. "Family Selection of Child Care Centers: The Influence of Household Support, Ethnicity, and Parental Practices." *Child Development* 67 (6): 3320–37. https://doi.org/10.2307/1131789.

Gelfand, Sharla. 2022. *Opendatatoronto: Access the City of Toronto Open Data Portal*. https://CRAN.R-project.org/package=opendatatoronto.

Hansmann, Henry. 1980. "The Role of Nonprofit Enterprise." *Yale Law Journal* 89 (5): 835–901.

Hansmann, Henry B. 1979. "The Role of Nonprofit Enterprise." *Yale LJ* 89: 835.

Harrell, Frank E. 2012. "Regression Modeling Strategies." *R Package Version*, 6–2.

Harrell, Frank E. 2015. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer.

Hlavac, Marek. 2022. *stargazer: Well-Formatted Regression and Summary Statistics Tables*. Bratislava, Slovakia: Social Policy Institute. https://CRAN.R-project.org/package=stargazer.

Hosmer, David W., Stanley Lemeshow, and Rodney X. Sturdivant. 2013. *Applied Logistic Regression*. Wiley.

Johnson, Anna D., Rebecca M. Ryan, and Jeanne Brooks-Gunn. 2012. "Child-care Subsidies: Do They Impact the Quality of Care Children Experience?" *Child Development* 83 (4): 1444–61. https://doi.org/10.1111/j.1467-8624.2012.1780.x.

Kaga, Yoshie, John Bennett, and Peter Moss. 2010. *Caring and Learning Together: A Cross-National Study on the Integration of Early Childhood Care and Education Within Education*. UNESCO.

Kuhn, and Max. 2008. "Building Predictive Models in r Using the Caret Package." *Journal of Statistical Software* 28 (5): 1–26. https://doi.org/10.18637/jss.v028.i05.

Larsen, John, Ahmed El-Geneidy, and Farzana Yasmin. 2015. "The Accessibility of Child Care Services: An Analysis of Toronto." *Journal of Transport Geography* 48: 41–49. https://doi.org/10.1016/j.jtrangeo.2015.08.005.

Morrissey, Taryn W. 2010. "Child Care and Child Development: What We Know and Why We Need to Know More." *Child Development Perspectives* 4 (2): 87–92. https://doi.org/10.1111/j.1750-8606.2010.00123.x.

Peng, Chao-Ying Joanne, Kuk Lida Lee, and Gary M Ingersoll. 2002. "An Introduction to Logistic Regression Analysis and Reporting." *The Journal of Educational Research* 96 (1): 3–14.

Penn, Helen. 2011. *Quality in Early Childhood Services: An International Perspective.* McGraw-Hill Education.

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.

Salamon, Lester M. 1995. *Partners in Public Service: Government-Nonprofit Relations in the Modern Welfare State.* Johns Hopkins University Press.

Singer, Judith D., and John B. Willett. 2003. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence.* Oxford University Press.

Steyerberg, Ewout W, Andrew J Vickers, Nancy R Cook, Thomas Gerds, Mithat Gonen, Nancy Obuchowski, Michael J Pencina, and Michael W Kattan. 2010. "Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures." *Epidemiology* 21 (1): 128–38.

Tekin, Erdal. 2007. "Child Care Subsidies, Wages, and Employment of Single Mothers." *Journal of Human Resources* 42 (2): 453–87. https://doi.org/10.3368/jhr.XLII.2.453.

Vines, Shere'lle Ramsey. 2020. "Accessing Equity: Challenges Middle-Income Families Face Finding High Quality Childcare."

Weisbrod, Burton A. 2000. *To Profit or Not to Profit: The Commercial Transformation of the Nonprofit Sector.* Cambridge University Press.

Weisbrod, Burton A. 1998. *To Profit or Not to Profit: The Commercial Transformation of the Nonprofit Sector.* Cambridge University Press.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. https://ggplot2.tidyverse.org.

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.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *dplyr: A Grammar of Data Manipulation.* https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *readr: Read Rectangular Text Data.* https://CRAN.R-project.org/package=readr.

Xie, Yihui. 2021. *knitr: A General-Purpose Package for Dynamic Report Generation in r.*

https://yihui.org/knitr/.