Factors of First-Year Financial Savings in Solar Photovoltaic Systems: Data from Open Data Toronto and SolarOT Map*

Finding Factors Effect First Year Bill Savings: Multiple Linear Regression Model

Yuanchen Miao

December 14, 2024

This paper examines the factors affect first-year financial savings for solar photovoltaic (PV) systems using data from Open Data Toronto and the SolarOT Map application. A multiple linear regression model identifies how variables such as payback period, roof size, system cost, electricity generation, and greenhouse gas (GHG) reduction influence initial bill savings. The analysis finds that payback period and system cost are significant predictors of first-year savings, while electricity generation has small influence and GHG reduction and roof size are not. These findings clarify the financial and physical conditions that maximize early solar benefits that can help improve system design, inform policy, and accelerate the adoption of renewable energy technologies.

Table of contents

*Code and data are available at: https://github.com/NevaeH-9/SolarTO.

1 Introduction

This study investigates the relationships between several key variables influencing the adoption of solar photovoltaic (PV) systems, utilizing data sourced from **Open Data Toronto** (City of Toronto 2024) and the SolarOT Map application. A multiple linear regression model was developed to analyze the impact of predictors such as payback period, roof size800k, system cost, annual electricity generation, and annualGHG reduction on the response variable, first year bill savings. The research aims to identify the most significant factors affecting early financial benefits from solar PV installations, providing some ideas to people on deciding whether they choose to have a solar PV system.

The primary estimand of this study is the average effect of predictors payback period, roof size, system cost, annual electricity generation, and annual GHG reduction on the expected first year bill savings. The model seeks to quantify how changes in these predictors influence the immediate financial benefits homeowners can expect during the first year of solar PV operation, adjusting for the unique characteristics of individual installations.

The multiple linear regression model reveals that payback period and system cost are statistically significant predictors of first year bill savings. In contrast, annual electricity generation shows a small influence on response variable and roof size and annual GHG reduction do not show a significant relationship with the response variable. This indicates that, while system cost and payback period is very important in determining initial financial benefits, the overall

electricity generation capacity and greenhouse gas reductions may have less direct influence on first year bill savings.

Understanding the factors that affects the early financial benefits is important for people who decides to have a solar PV system, as many potential users prioritize immediate cost savings when making investment decisions. By highlighting the significance of variables like payback period, annual electricity generation, and system cost, this paper provides idea for policymakers, manufacturers, and installers aiming to design cost-effective solar PV systems. Furthermore, the findings underscore the importance of tailoring incentives and educational efforts to emphasize the factors that most directly impact early financial returns, potentially accelerating the transition to sustainable energy solutions.

The rest of paper are structured as follows: Section 2 introduced the programming language that is used in this paper and packages are used to generate models, graphs and tables, and tidy the paper. The predictors are also introduced in the section. Section 3 is introducing the multiple linear regression model, the coefficients of each predictor and the model validation and justification. The result of the model is discussed in Section 4. Overview of the paper, how first year bill saving encourage people to install solar system, why is solar PV system better than regular electricity generation and weakness and nextstep for this paper are discussed in Section 5.

2 Data

2.1 Overview

The statistical programming language **R** (R Core Team 2023) is used as a base of this paper. Packages **lubridate** (Grolemund and Wickham 2011), **dplyr** (Wickham et al. 2023), **tidyverse** (Wickham et al. 2019) and **opendatatoronto** (Gelfand 2022) are used to download, clean the data and generate the multiple linear regression model. Packages **ggplot2** (Wickham 2016) and **gridExtra** (Auguie 2017) has been used to make figures and labeling the graph. Packages **kableExtra** (Zhu 2024) and **broom** (Bolker and Robinson 2024) have been used to generate tables for the paper. Package **modelsummary** (Arel-Bundock 2022) has been used to summary the model. Package **lmtest** (Zeileis and Hothorn 2002) has been used to provide test for model validation. Package **arrow** (Richardson et al. 2024) has been used to save the cleaned data in parquet form. The data was downloaded from Open Data Toronto that refreshed daily and collected using SolarTO Map. With the cleaning process, observations with NA values and address that is not found by the software are removed from the data set. Multiple linear regression model are being used to find the relationship between first year bill savings with other factors.

2.2 Clean Data

Some data points in the variables system size, system cost, annual electricity generation, and annual GHG reduction were identified as significantly higher than the majority of the recorded values. These data points were classified as outliers and excluded from the dataset to ensure a more robust analysis. To investigate the relationship between first-year bill savings and other predictors, only observations with values below the mean for system size, system cost, annual electricity generation, and annual GHG reduction were retained. The original dataset contained 522,428 observations, which were reduced to a subset of 5,000 observations to optimize computational efficiency during model training. Additionally, observations where system size equals zero were removed from the dataset. A system size of zero indicates that the roof of the building is not suitable for installing a solar PV system. As these observations do not contribute to identifying the relationship between the predictors and the response variable, they were excluded from the analysis.

2.3 Measurement

The data set is downloaded from **Open Data Toronto** (City of Toronto 2024), a website that post data collected in Toronto City can be used and re-published as everyone wish freely that the website is aiming to make data innovated by anyone to draw insights and use evidence to inform the design of civic services. The raw data is collected by using SolarTO Map, a Geographic Information Systems analysis of Light Detection and Ranging data. The software takes into account geographical latitude and the sun's daily position throughout the year. By using SolarTO Map, people are able to find out that the data of the roof area they could put a device and how much solar energy it could generate and how much money they are saving by using the solar energy device by selecting the area they are looking for. First several rows of cleaned data is in Table 1

Table 1: First Several Rows of Cleaned SolarTO Map Data provided by Open Data Toronto

ID	Elec.Gen.	Bill Savings	System Size	Payback	GHG Reduc.	Roof Size	System Cost
38 Westleigh Cres Structure 1	5627	881	4.893044	10	394	44	14700
36 Westleigh Cres Structure 1	6012	941	5.227826	9	421	51	14100
34 Westleigh Cres Structure 1	4061	636	3.531304	10	284	32	10600
34 Westleigh Cres Structure 2	1480	232	1.286957	10	104	12	3900

2.4 Outcome variables

The outcome variable in this paper is first year bill savings that it represents the estimated electricity bill savings over the first year of operation of the solar PV system in dollars. This is calculated by multiplying the average utility rate with the amount of electricity produced by solar panel but not consumed by users. First year bill savings is also considered as response variable in the multiple linear regression model.

2.5 Predictor variables

Annual Electricity Generation(annual_electricity_generation_k): Estimated annual electricity production of the proposed rooftop solar system measured in kWh. To estimate Annual Electricity Generation the following assumptions are used: 15% panel efficiency and 86% performance ratio.

System Size(system_size): The size of the solar photovoltaic (PV) system that can fit on the rooftop, measured in kilowatts (kW). It is assumed that 1 kW of solar will generate 1,150 kWh/year, based on average solar radiation in Toronto. The System Size is calculated by dividing the Annual Electricity Generation (kwh) by 1,150 kWh.

Payback Period(payback_period): The payback period refers to the number of years it will take to recover the cost of the solar PV system through the savings generated by the solar PV system. The payback period is based on annual electricity bill savings minus the upfront cost of the solar system.

Annual GHG Reduction(annual_ghg_reduction_kg): Greenhouse gas reduction is based on the reduced consumption from the grid and it's associated GHGs each year, measured in kilograms of carbon dioxide equivalent (kg of CO2e). It is assumed that solar will offset 0.07 kg (000.07 tonnes) of CO2e (carbon dioxide equivalent) for each kWh of solar electricity produced.

*Roof Size(roof_size800k)**: Total roof area that receives at least 800 kWh of (kilowatt hours) of solar radiation per square metre. Based on industry best practices, the SolarTO Map uses certain criteria to identify a suitable rooftop, one of which is that the rooftop receive at least receives at least 800 kWh of solar radiation per square meter.

System Cost(system_cost): The upfront cost of the solar PV system in dollars. The System Cost is based on the System Size and the average solar installation rates in Toronto. A estimated $/Watt rate ranging from $3/watt to $2,2/watt (dependent on system size) is applied to calculate cost.

Figure 1 shows that distribution of system cost, the majority of the system cost is around 15000 dollars, and minority of system is cost more than 27000. The graph has no general trend and the price of any system is possible. The system cost exhibits randomness.

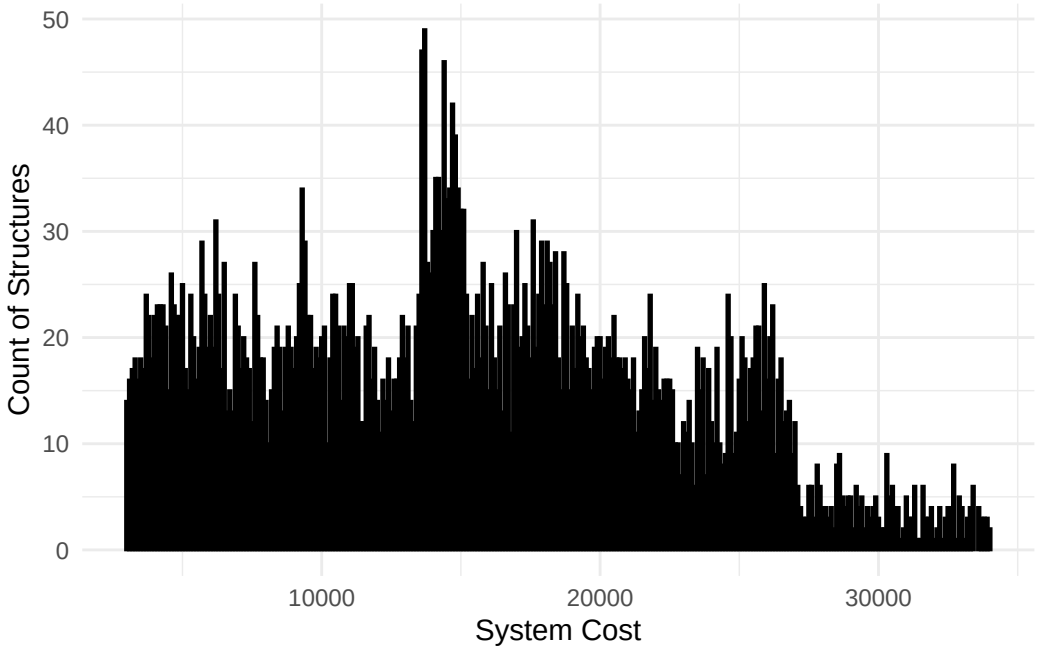

Figure 1: Distribution of System Cost

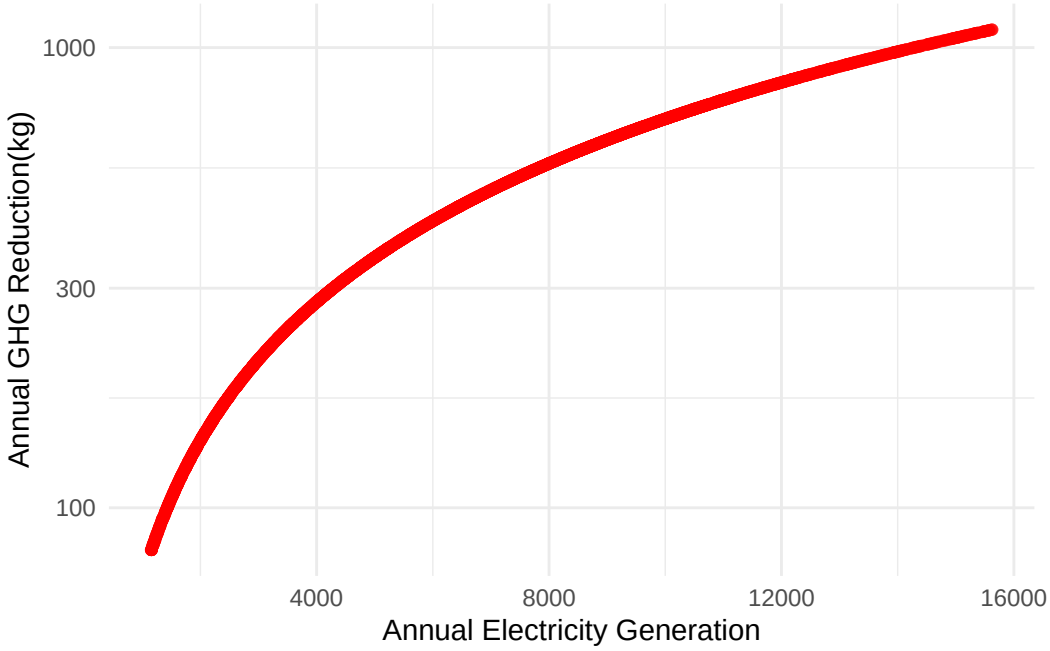

Figure 2: Distribution of System Size and Annual GHG reduction

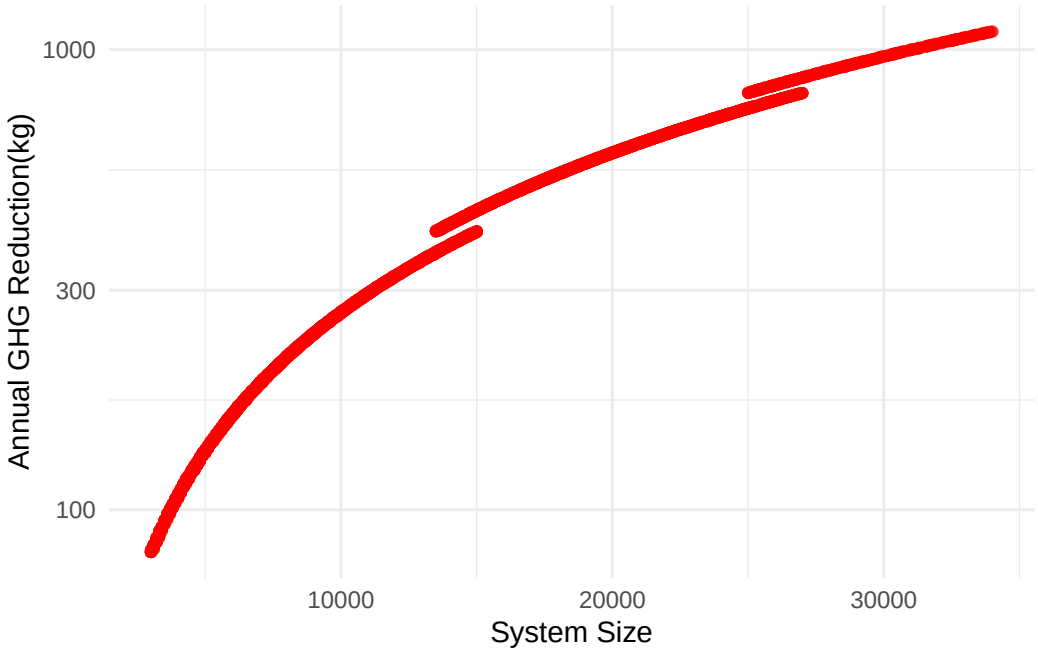

Figure 3: Distribution of System Size and Annual GHG reduction

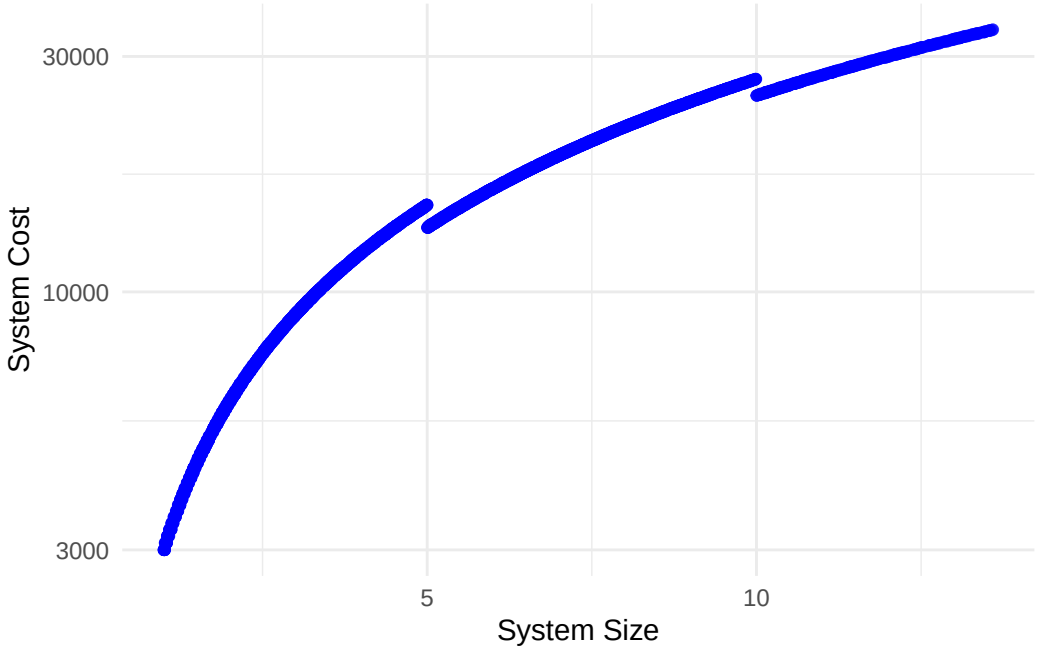

Figure 4: Distribution of System Size and System Cost

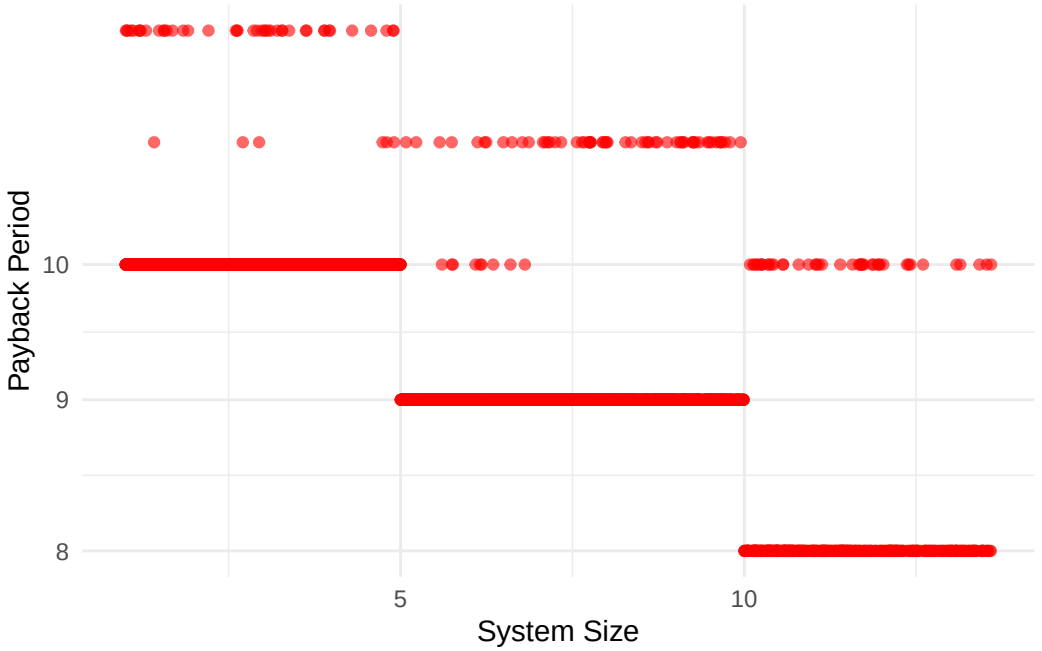

Figure 5: Distribution of System Size and Payback Period

The roof size refers to the total area capable of receiving at least 800 kWh of solar radiation per square meter. To maximize the efficiency of the solar photovoltaic (PV) system, it is assumed that the entire roof area will be utilized for system installation. It is further assumed that as the roof size increases, the system size will also increase proportionally. As demonstrated in Figure 2 and Figure 3, there is a clear positive relationship between system size, annual electricity generation, and annual greenhouse gas (GHG) reduction. Specifically, larger PV systems generate more electricity than smaller systems, leading to a reduction in electricity consumption from alternative sources. Consequently, this results in a greater annual GHG reduction for larger systems.

In Figure 4 and Figure 5, it is observed that the payback period is not directly correlated with system size. While the cost of larger systems is higher, these systems also generate more electricity, leading to greater savings on electricity bills. The number of years required to recover the initial investment through bill savings is also related to the system size, even payback period differed from same system size, but larger system size is having a trend to have shorter payback period.

3 Model

The multiple linear regression model is generated to investigate the relationships between annual billing savings and predictor system cost, roof size and payback period. These predictors are selected for significance in influencing the change of the annual billing savings and the other predictors are reduced by insignificance.

Here we briefly describe the multiple linear regression model used to investigate the relationship.

3.1 Model set-up

$$\text{Annual Billing Savings} = \beta_0 + \beta_1 \cdot \text{system size} + \beta_2 \cdot \text{payback period} + \beta_3 \cdot \text{system cost}$$
$$+\beta_4 \cdot \text{annual electricity generation} + \beta_5 \cdot \text{annual GHG reduction}$$

- β_0 is the intercept of the multiple linear regression model, which represents the value of annual billing savings when other predictors are zero
- β_1 is the coefficient of **system size** that represents the effect of adding system size on annual billing savings
- β_2 is the coefficient of **payback period** that represents the effect of adding a year on payback period on annual billing savings
- β_3 is the coefficient of **system cost** that represents how much of adding a a dollar has effect on system cost on annual billing savings
- β_4 is the coefficient of **annual electricity generation** that represents the change of annual billing savings when one kWh electricity is generated by the solar PV system
- β_5 is the coefficient of **annual ghg reduction** that represents the change of annual billing savings when one kg carbon dioxide is reducted by using the solar PV system

3.2 Model Justification

In this paper, a multiple linear regression model is used to predict first year bill savings. Payback period is included as a predictor because it reflects the time required to recover the investment in the solar system, which is an essential consideration for prospective solar adopters. This variable is expected to have a direct relationship with bill savings, as shorter payback periods generally correlate with higher savings. Roof size and system size are selected from raw data to be predictors since a larger roof area allows for the installation of a more extensive solar system, and larger system size. Both of these predictors indicate more electricity is generated from the system, leads to higher billing savings. System cost and annual electricity generation are included as they directly influence the savings generated, with a more expensive system generally producing more electricity and offering a greater savings. Annual GHG reduction is also relevant since it is a proxy for the environment impact of a solar system,

with larger reductions correlating with more substantial savings due to reduced reliance on non-renewable energy sources.

3.3 Model Assumptions and Validation

- **Linearity**: Assumes a linear relationship between predictors and the response variable.
- **Homoscedasticity**: Residuals are having constant variance.
- **Independence**: Residuals are independent of each other.
- **Normality of Errors**: Residuals are normally distributed

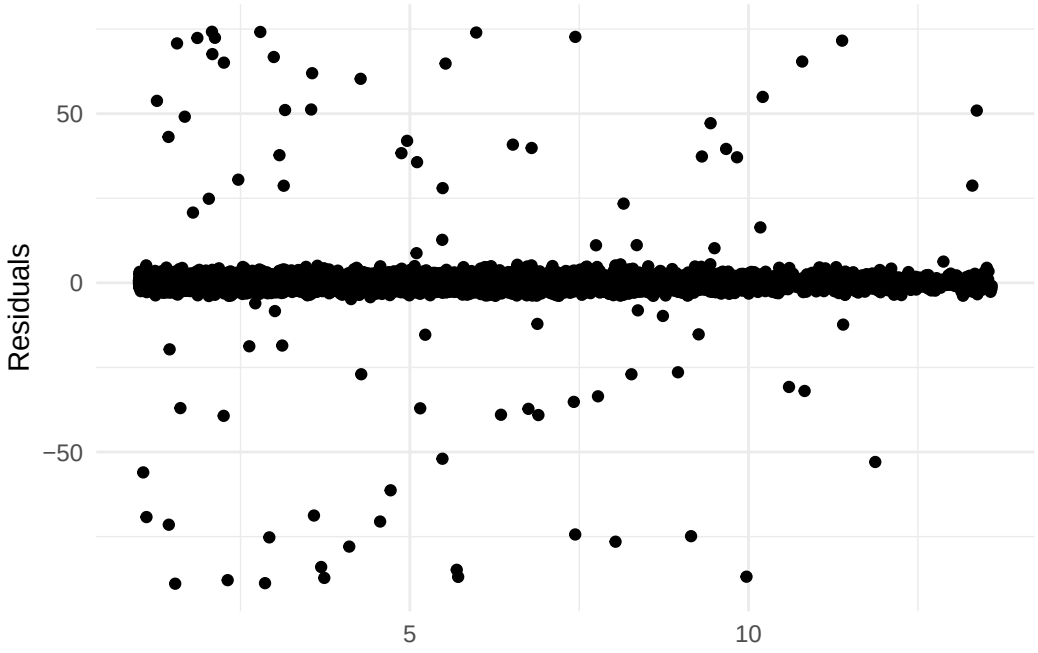

Figure 6: Residual Plot of Multiple Linear Regression Model

In Figure 6, we can inspect the relationship between the residuals and fitted values. The scatter plot of residuals versus fitted value shows no pattern on the graph but it has a lot of outliers, the model is violating the assumption on linearity.

Table 2: Breusch-Pagan Test for Homoscedasticity

	Statistic	p-value
BP	2696.866	0

Homoscedaticity assumes constant variance of residuals across all levels of the independent variables. The plot of residuals versus fitted values shows the identity but we can further use the Breusch-Pagan test to check for homoscedasticity. The null hypothesis of the Breusch-Pagan test is that there is constant variance. A p-value above 0.05 indicates that homoscedasticity holds. From Table 2, it shows that we are having the p-value as zero which indicates that the model fails on the homoscedasticity test.

Table 3: Durbin-Watson Test for Autocorrelation

	Statistic	p-value
DW	1.752778	0.7603159

To check for independence of residuals, I am using the Durbin-Watson test for autocorrelation, which tests whether residuals are correlated. A p-value greater than 0.05 suggests that the residuals are independent. From Table 3, we are having a p-value 0.76 which is greater than 0.05, thus, residuals are independent.

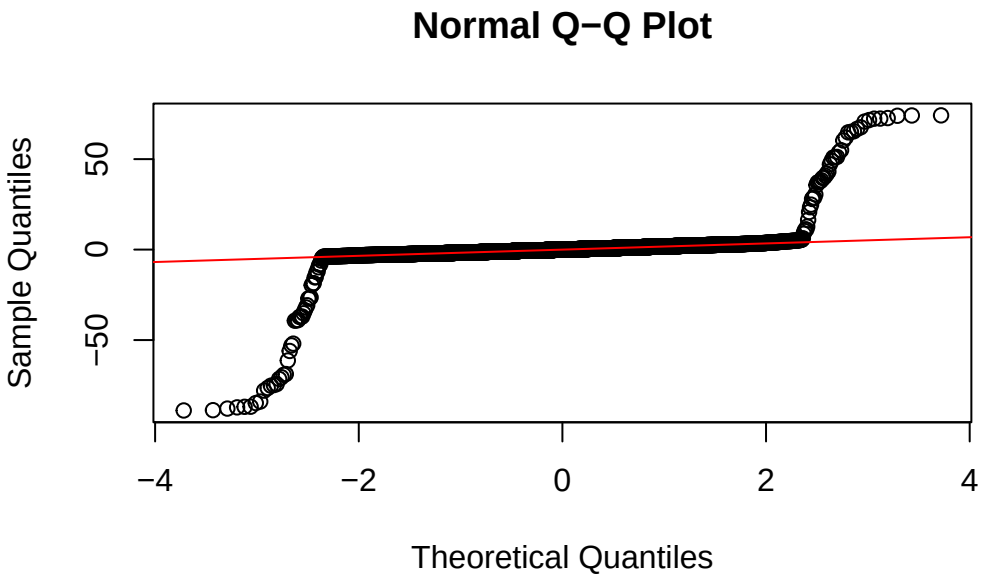

Figure 7: Histogram and Q-Q Plot of Multiple Linear Regression Model

To check if the residuals are normally distributed, we are using the Q-Q plot and histogram of residuals. If the residuals follow a normal distribution, the points in the Q-Q plot should lie along a straight line, and the histogram should resemble a bell shaped curve. In Figure 7, the

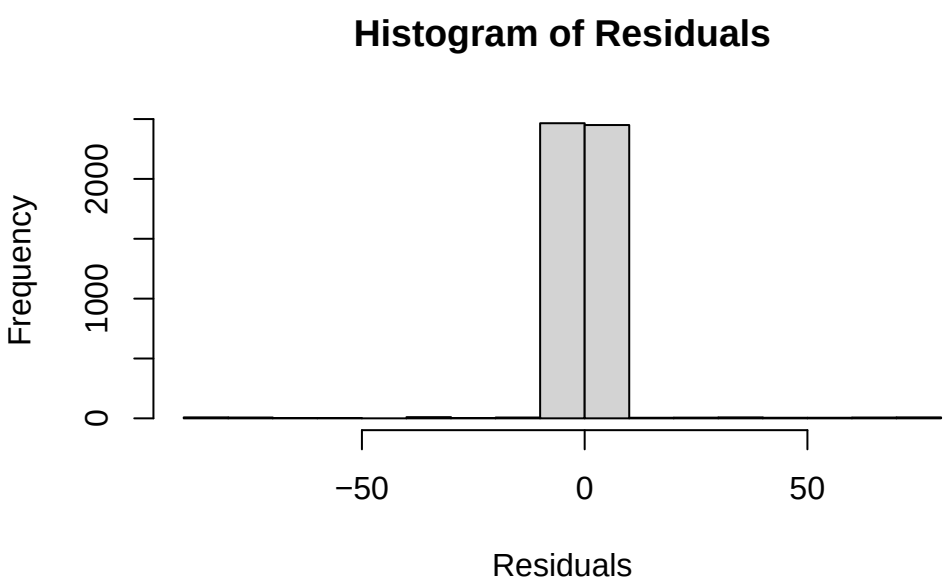

Figure 8: Histogram and Q-Q Plot of Multiple Linear Regression Model

Q-Q plot is not a straight line and in Figure 8 the histogram is showing as a rectangle shape. These two graphs indicates that the model is not normally distributed.

4 Result

Table 4: Model Summary

term	estimate	std.error	statistic	p.value
(Intercept)	575.0139095	3.8349491	149.9404275	0.0000000
payback_period	-57.2082431	0.3798780	-150.5963399	0.0000000
roof_size800k	-0.0577577	0.0438878	-1.3160305	0.1882242
system_cost	0.0358037	0.0003406	105.1323259	0.0000000
annual_electricity_generation_k	0.0529565	0.0249405	2.1233101	0.0337769
system_size	NA	NA	NA	NA
annual_ghg_reduction_kg	0.1394592	0.3560014	0.3917377	0.6952687

As shown in Table 4 the summary of the model, the predictor payback period has a $p-value < 2 \times 10^{-16}$ and an estimate -57.2082431 indicates that it has a strong negative relationship

suggests shorter payback periods are associated with greater first-year billing savings. System cost also has a $p-value < 2 \times 10^{-16}$ and has estimate equals to 0.0358037 which indicates that it has a positive relationship with the response variable, higher system cost tend to generate more first-year savings.

Annual electricity generation has a p-value $0.0338 < 0.05$ that it has a weaker positive association suggests that the more electricity the system saves, more first year bill saving is obtained.

Predictors roof size ($p - value = 0.1882$) and annual GHG reduction ($p - value = 0.6953$) are both having $p - value > 0.05$. These two predictors are not strongly associated with the response variable in this dataset.

Predictor system size is showing NA in the summary of the model that it is not estimable due to singularities. System size is dependent with one of other predictors, thus, system size should be removed as a predictor in the model.

5 Discussion

5.1 Overview of Paper

In this paper, I investigates the relationship between various factors affecting the first-year bill savings from solar photovoltaic (PV) systems. Using data collected from the SolarOT Map app and Open Data Toronto, the paper applies a multiple linear regression (MLR) model to predict the first-year bill savings based on several key predictors. These predictors include payback period, roof size, system cost, annual electricity generation, and annual GHG reduction. The model is carefully constructed and validated to identify which factors have the most significant influence on the savings achieved by installing a solar PV system. Significance of each predictor are explored through statistical analysis and model diagnostics, including tests for linearity, homoscedasticity, normality of residuals and independence.

5.2 How First Year Bill Saving Encourage People to Install Solar PV System

First year bill savings are a key factor that can motivate individuals to install solar PV systems. These savings directly impact the financial benefits of adopting solar energy by lowering electricity costs, making the initial investment more attractive. Many homeowners and businesses are attracted to the potential for immediate savings on their energy bills. When they see a reduction in they utility costs in the first year of installation, it provides a quick return on investment, making the decision to install solar more appealing. This early financial benefit can offset the initial costs associated with purchasing and installing a solar PV system..

The first year savings help to speed up the recovery of the upfront cost of the solar system. If the savings in the first year are significant enough, it can shorten the payback period. The faster people can recoup their investment, the more likely they are to proceed with installation.

The promise of tangible first year saving can lower the perceived financial risk of investing in solar energy. Potential adopters may feel more secure in their decision if they can expect visible financial returns. This can be especially important for people who are cautious about making large investments in new technologies.

5.3 Solar PV System versus Regular Electricity Generator

Solar PV systems have become increasingly advantageous in today's world due to several key factors. Solar PV systems help reduce reliance on fossil fuels, which are the primary contributors to greenhouse gas emissions. By generating electricity by solar energy, it is changing and reducing global warming potential. Unlike fossil fuels, solar energy is inexhaustible. The sun provides an enormous amount of energy, and harnessing it through solar PV systems does not deplete natural resources, making it a sustainable solution for the long term.

The cost of installing solar PV systems has decreased significantly over the past few years due to technological advancements, economies of scale, and government incentives. This makes solar energy more affordable and accessible to a wider range of people and businesses. Solar PV systems can significantly reduce electricity bills by generating energy on-site. This leads to lower utility costs and can provide long-term savings. Additionally, surplus energy can often be sold back to the grid, further reducing cost or even generating income.

Solar PV systems allow homes and business to generate their own electricity, reducing reliance on the public power grid. This is beneficial during power outages or times of high demand when grid stability maybe at risk. By integrating solar energy with battery storage solutions, users can ensure a more reliable and self-sufficient energy supply, even during grid failures. This enhances energy security and resilience, particularly in remote areas or regions prone to natural disasters.

Solar PV system also have a lifespan of 25-30 years or more with minimal maintenance. This makes them a long-term investment that can continue to generate savings over time, especially as utility prices rise. Homes with solar panels are often valued higher than those without, as potential buyers recognize the value of reduced energy bills and energy independence. Solar PV systems can contribute to increasing property value.

5.4 Weaknesses and Next Steps

In the model, system size is a predictor I considered it will have influence on first year bill saving but it is correlated with other predictors that it shows NA on model summary.

For future research, I would investigate on which building has solar PV system installed. By having data of buildings with solar PV system and combine with the data I obtained from Open Data Toronto, it is possible to build a Bayesian regression model to predict buildings people are more willing to install a solar PV system.

A Appendix

B Additional data details

Table 5: Summary Table of Raw Data Columns

Column Name	Description
structureid	Unique identifier for the structure (e.g., a building or house).
annual_electricity_generation_k	The annual electricity generated by the solar PV system, in kilowatt-hours (kWh).
first_year_bill_savings	The savings on electricity bills in the first year of solar PV system operation.
system_size	The size of the solar PV system, typically measured in kW (kilowatts).
payback_period	The number of years required to recover the cost of the solar PV system through savings.
annual_ghg_reduction_kg	The amount of greenhouse gas (GHG) emissions reduced annually by the solar PV system, in kilograms (kg).
roof_size800k	The total roof area of the building that can receive at least 800 kWh of solar radiation per square meter.
system_cost	The total cost of installing the solar PV system.

C Methodology of SolarTO Map Calculating Data

Input Data

- **Geospatial Data**: Roof geometry, slope, and orientation are derived from high-resolution satellite imagery and LiDAR data. These datasets enable accurate modeling of the roof area and its suitability for solar panels.
- **Solar Radiation Data**: Hourly solar irradiance values are collected from historical weather datasets. These measurements are adjusted for local conditions such as cloud cover, seasonal variations, and shading.
- **Shading Analysis**: The tool employs algorithms to estimate shading from nearby structures, trees, and other obstructions. Shading reduces solar panel efficiency and affects electricity generation estimates.

Annual Electricity Generation:

The annual electricity generation E_{gen} is estimated using the formula:

$$E_{\text{gen}} = A \cdot G \cdot \eta$$

Where: - $A$is the usable roof area m^2. - G is the solar irradiance kWh/m^2/year. - η is the solar panel efficiency.

System Cost: The estimated system cost is calculated by multiplying the system size by the cost per kilowatt. The cost per kilowatt is derived from market averages and may include incentives or rebates offered by government programs.

Payback Period:

$$\text{Payback Period} = \frac{\text{Total System Cost}}{\text{Annual Savings}}$$

Where annual savings include reductions in electricity bills due to solar energy generation. The tool assumes average electricity prices and utility rate structures.

First-Year Bill Savings: The first-year bill savings are calculated based on the amount of electricity generated in the first year and the average cost of electricity. Adjustments are made for seasonal variations in energy consumption and local utility rates.

Annual GHG Reduction:The annual GHG reduction is calculated by estimating the emissions avoided when electricity from solar replaces electricity from fossil fuels.

$$\text{GHG Reduction} = E_{\text{gen}} \cdot E_{\text{factor}}$$

Where: - E_{gen} is the annual electricity generation. - E_{factor} is the emission factor, reresenting the amount of carbon dioxide avoided per kWh of electricity granted by solar PV system.

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Auguie, Baptiste. 2017. *gridExtra: Miscellaneous Functions for "Grid" Graphics*. https://CRAN.R-project.org/package=gridExtra.

Bolker, Ben, and David Robinson. 2024. *Broom.mixed: Tidying Methods for Mixed Models*. https://CRAN.R-project.org/package=broom.mixed.

City of Toronto. 2024. "SolarTO Map Dataset." https://open.toronto.ca/dataset/solarto/.

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

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

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.

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.

Zeileis, Achim, and Torsten Hothorn. 2002. "Diagnostic Checking in Regression Relationships." *R News* 2 (3): 7–10. https://CRAN.R-project.org/doc/Rnews/.

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.