Debt as a socioeconomic determinant of cancer incidence

Fysal Beauferris - 30048436

1. INTRODUCTION

1.1. MOTIVATION

1.1.1. Context

Cancer is a complex and multifactorial disease that can be caused by a variety of factors, including genetic mutations, environmental factors, lifestyle choices, and certain infections. **Cancer is the number 1 cause of death in Canada, 2 in 5 Canadians will develop cancer in their lifetime and 1 in 4 Canadians will die from cancer [11].** Despite significant advances in cancer research, the exact causes of many types of cancer are still not fully understood. Thus, identifying more causes of cancer is important because it can help us better understand the disease and develop more effective prevention and treatment strategies. To gain a comprehensive understanding of cancer, it is important to take a holistic approach that examines the potential social and economic factors that may contribute to its development, rather than solely focusing on its preestablished and agreed upon causes.

The project at hand can be applied to various key domains of cancer research, such as prevention and screening, diagnosis and prognosis, treatment, survivorship and supportive care, and health disparities. These applied domains aim to identify risk factors, develop better diagnostic and treatment options, improve the quality of life for cancer survivors, and address social, economic, and cultural factors that contribute to disparities in cancer outcomes.

1.1.2. Problem

Current research suggests a possible link between mental health conditions such as chronic stress, depression, or anxiety and cancer. Additionally, research suggests increasing debt may lead to increased stress/depression/anxiety in individuals. Therefore, the objective of this study is to examine the relationship between socioeconomic factors such as debt and stress and investigate whether there is a direct correlation between these factors and cancer occurrence. Specifically, the study aims to analyze the extent to which debt and mental health conditions are associated with cancer incidence and explore potential causal mechanisms underlying any observed relationships.

1.1.3. Challenges

Given that the aim of this study is to investigate whether there is a direct correlation between socioeconomic factors such as debt and stress, and cancer occurrence, it is essential to consider that correlation does not necessarily indicate causation. There may be speculation among people regarding the connection between cancer and non-physical health factors. Therefore, this study will explore and demonstrate how factors like debt can cause stress, which can weaken an individual's immune system, making them more susceptible to a cancer diagnosis. To broaden the focus beyond one-to-one relationships, the study will explore more path-based relationships in data that can contribute to cancer diagnosis.

One of the challenges in this study is the limited existing literature indicating a connection between debt and cancer incidence. The lack of related research does not necessarily mean that the relationship is invalid, but it highlights the overemphasis on establishing one-to-one causal links in cancer causes, rather than exploring more complex path-based relationships. Therefore, the study's task is to establish a path that demonstrates how debt can be linked to cancer incidence via intermediate factors such as stress.

1.2. OBJECTIVES

1.2.1. Overview

There is a lack of research on the relationship between debt and cancer incidence, which highlights the need to explore underexplored socioeconomic factors that may contribute to an increase in cancer incidence.

During the period around 2008, there was a significant recession that impacted employment and unemployment rates in Canada. In October 2009, employment in Canada was **down 400,000** from the peak in October 2008, **a loss of 2.3%** in seasonally adjusted figures, and during the same period, the unemployment rate rose from **6.3% to 8.6%** (LaRochelle-Cote et al., 2009, p.1) **[5]**. Identifying a correlation between debt and cancer incidence during this time could potentially link this trend to the 2008 recession and serve as an indicator or predictor of a similar outcome in the near future, particularly in light of ongoing speculation about the possibility of another recession in Canada.

The aim of exploring the potential relationship between debt and cancer is to prompt government-funded health institutions and other large organizations to allocate more resources to raising awareness of non-physical factors that may

contribute to cancer incidence. By exploring how debt can impact physical health and increase susceptibility to cancer, there could be a greater understanding of how debt can affect cancer incidence and potentially relate it to cancer remission.
Allocating resources towards proactive measures for addressing cancer, rather than solely directing billions of taxpayer dollars towards treating cancer patients at homes and in hospitals, could be a more effective approach. For instance, if this research project establishes a sound relationship between debt and cancer incidence, the government of Canada could consider putting additional funding towards tackling the debt crisis in Canada. This approach could potentially lower the rate of cancer incidence.
From a societal perspective, cancer-related costs were **CAD 26.2 billion in Canada in 2021**, with **30%** of costs being borne by patients and their families. The economic burden was the highest in the first year after cancer was diagnosed (i.e., initial care), with patients and families' costs amounting to almost **CAD 4.8 billion in 2021** (Garaszczuk et al., 2022, p.2745) **[9].**

1.2.2. Goals & Research Questions

- **Q1.** Do different trends or patterns exist between debt and cancer incidence when changing gender/age/geography in cancer incidence population?
  - **Goal 1:** Discover a stronger relationship than already established when investigating between debt and cancer incidence when changing gender/age/geography in cancer incidence population.
- **Q2.** Is there a trend around the years surrounding 2008? If so, can we relate it to the 2008 recession in Canada?
  - **Goal 2:** Discover a trend around the years surrounding 2008 as well as finding supporting evidence for the increase in household debt in Canada due to the 2008 recession.
- **Q3.** If we can identify a trend with the 2008 recession, considering the current speculation of the Canadian economy entering a new recession, can we predict a potential uptick in cancer incidence in that future scenario?
  - **Goal 3:** Create a simple linear regression model on the relationship and use R to predict changes in cancer incidence (dependent variable) with changes to debt (independent variable).
- **Q4.** If we can establish a relationship between debt and cancer incidence, can we extend this to cancer remission? Is debt related to cancer treatment?
  - **Goal 4:** Find supporting academic papers for my hypothesis that they are related.
- **Q5.** If we can establish a relationship between socioeconomic factors such as debt and stress and investigate whether there is a direct correlation between these factors and cancer occurrence?
  - **Goal 5:** Create a multiple regression model on the relationship and use R to predict changes in cancer incidence (dependent variable) with changes to debt (independent variable #1) and stress (independent variable #2).

1.3 METHODOLOGY

1.3.1. Data

1. Title: Number and rates of new cases of primary cancer, by cancer type, age group and sex.
Source: Statistics Canada. (2022). Table 13-10-0110-01 *Number and rates of new cases of primary cancer, by cancer type, age group and sex.*
Retrieved from https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310011101
Data Collection Method: Data collected from the Canadian Cancer Registry (CCR) which is a population-based registry that includes data collected and reported to Statistics Canada by each provincial/territorial cancer registry.
Attributes:

- **Geography:** Indicates the location of cancer cases, including Canada as a whole, provinces and territories, and Canada excluding Quebec.
- **Age group:** Identifies the age groups of individuals affected by cancer, ranging from 0-4 years to 90 years and over, as well as an age group where age is not stated.
- **Sex:** Identifies whether the cancer cases reported are for both sexes, males or females.
- **Primary Types of Cancer:** Lists the primary sites of cancer, including different types of cancer that occur in various parts of the body.
- **Characteristics:** Includes the number of new cancer cases, **cancer incidence** rate, and low and high 95% confidence intervals for the incidence rate.
- **Reference Period:** 1993- 2021

Cancer incidence refers to the number of new cases of cancer that are diagnosed during a specific time, typically expressed as a rate per 100,000 people per year. It is an important measure of the burden of cancer in each population and is used to track trends in cancer rates over time, as well as to identify patterns in the distribution of cancer by age, sex, race, and other demographic factors. Cancer incidence can be calculated for specific types of cancer (e.g., breast cancer, lung cancer, etc.) or for all types of cancer combined.

2. Title: Household sector credit market summary table, seasonally adjusted estimates
Source: Statistics Canada. (2023). Table: 38-10-0238-01 *Household sector credit market summary table, seasonally adjusted estimates.*
Retrieved from https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3810023801
Data Collection Method: The National Balance Sheet Accounts (NBSA) record the stock of assets (financial and non-financial), liabilities and net worth for each institutional sector.
Attributes:

- **Geography:** Canada, refers to the entire country.
- **Categories:** All categories are Seasonally adjusted at quarterly rates.
  1. **Credit market debt to disposable income**
  2. **Consumer credit and mortgage liabilities to disposable income**
  3. **Mortgages - national balance sheet (x 1,000,000)**
- **Reference Period:** Q1 of 1990 - Q4 of 2022

Consumer credit and mortgage liabilities and credit market debt are related concepts that measure different aspects of household debt in Canada. ***Consumer credit and mortgage liabilities*** refer specifically to the amount of debt held by Canadian households for personal consumption and home ownership, respectively, relative to their disposable income. This measure includes loans and credit extended to individuals for personal use, such as credit card debt, personal loans, lines of credit, and mortgages on residential properties.

$$\textbf{Consumer Credit and Mortgage Liabilities to Disposable Income} = \frac{\textbf{Consumer Credit and Mortgage Liabilities}}{\textbf{Disposable Income}} \times \textbf{100}$$

Credit market debt, on the other hand, is a broader measure that includes all debt issued by the private sector in Canada, including households, businesses, and governments. This measure includes not only consumer credit and mortgage liabilities but also other forms of debt, such as corporate bonds, government bonds, and other loans.

$$\textbf{Credit} - \textbf{Market Debt to Disposable Income} = \frac{\textbf{Credit} - \textbf{Market Debt}}{\textbf{Disposable Income}} \times \textbf{100}$$

Mortgages - national balance sheet x 1000000 refers to the total value of outstanding residential mortgages on the national balance sheet in Canada, expressed in millions of dollars. This attribute captures the amount of debt that Canadian households have incurred through mortgages. The national balance sheet is a comprehensive accounting statement that measures the assets, liabilities, and net worth of the Canadian economy.

3. Title: Perceived life stress, by age group
Source: Statistics Canada. (2022). Table: 13-10-0096-04 *Perceived life stress, by age group.*
Retrieved from https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310009604
Data Collection Method: Data gathered by the Canadian Community Health Survey (CCHS), who's central objective is to gather health-related data at the sub-provincial levels of geography (health region or combined health regions).
Attributes:

- **Geography:** Indicates the locations data was collected, including Canada (excluding territories) and provinces.
- **Age group:** Identifies the age groups of individuals affected by stress, ranging from 12 years to 65 years and over.
- **Sex:** Identifies whether the cancer cases reported are for both sexes, males or females.
- **Indicators:** Population aged 12 and over who reported perceiving that most days in their life were quite a bit or extremely stressful. **Perceived life stress** refers to the amount of stress in the person's life, on most days, as perceived by the person or, in the case of proxy response, by the person responding.

- **Reference Period:** 2003-2021

1.3.2. Approach

I'll use R Studio to analyze and visualize data. To validate plot accuracy, I'll calculate summary statistics like the value of the test statistic and associated P-value and interpret them in the context of the plots. I'll also create linear and multiple regression models, including regression lines, labels to observation points, and include legends. Finally, I'll create confidence intervals for the linear regression models to provide a better understanding of the plot's range of values.

1.3.3. Workflow

To successfully complete my project, I must first create a task list which will consist of the following:

1. **Define the research question**
2. **Collect relevant data**
3. **Analyze the data**
4. **Interpret the results**
5. **Draw conclusions**

Establishing a causal relationship between debt and cancer incidence is a complex issue that requires careful consideration of other factors that could contribute to cancer development. In case our data collection or analysis is unsuccessful, alternative methods of analysis should be attempted, such as observing new datasets or subsets of our existing data. It is essential to note that correlation does not necessarily imply causation, and there could be other variables that are not accounted for in the analysis. Therefore, interpreting the results of any analysis requires careful consideration of alternative explanations for any observed relationships. The same level of scrutiny must be applied when examining the relationship between socioeconomic factors, such as debt, stress, and cancer occurrence using multiple linear regression models. It is crucial to consider the study's limitations and acknowledge that it may not capture all the factors that could affect the outcome.

1.3.4. Workload Distribution

As this is an independent research project, all the workloads will be my personal responsibility.

1.3. CONTRIBUTIONS

This study aims to investigate the relationship between credit market debt to disposable income and cancer incidence in Canada, as well as the relationship between consumer credit and mortgage liabilities to disposable income and cancer incidence in Canada. In addition, the study will explore the relationship between both mortgage loans and perceived life stress in relation to cancer incidence. While the relationship between debt and cancer has been lightly explored in literature, this research is unique as it analyses these specific debt statistics. The data sets used in the study are from Canadian sources, which further distinguishes this research as novel. The results of this study will provide valuable insights into the complex interplay between debt, stress, and cancer incidence, which could have important implications for public health policy and government based financial decision-making.

2. RELATED WORK

2.1. TECHNOLOGY SCAN

- **PRISMA:**

The acronym PRISMA stands for Preferred Reporting Items for Systematic Reviews and Meta-Analyses. It is a set of evidence-based guidelines that define the minimum items required for reporting systematic reviews and meta-analyses. The PRISMA statement includes a checklist of 27 items and a 4-phase flow diagram. These tools help ensure that systematic reviews and meta-analyses are reported in a clear, transparent, and standardized way. **[4]**

- **Meta-Analysis:**

Meta-analysis is a statistical technique used to combine and analyze data from multiple independent studies on a specific research question. The goal of a meta-analysis is to provide a more comprehensive and precise estimate of the true effect size of a particular intervention or phenomenon than any single study can provide.

By pooling data from multiple studies, meta-analysis can identify patterns and trends across studies, increase statistical power, and evaluate the consistency and heterogeneity of results. It can also identify potential biases in individual studies and assess the generalizability of findings across different populations, settings, and contexts. **[1]**

- **95% Confidence Intervals:**

Confidence intervals were calculated for the adjusted RR of cancer incidence, cancer-specific mortality, and all-cause mortality in cancer patients with depression and anxiety. [**7]**

- **Adjusted Relative Risk (RR):**

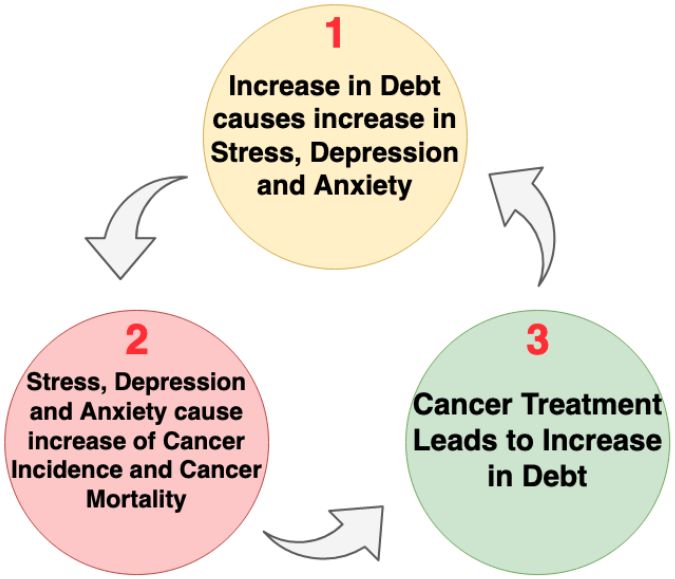
Relative risk is a statistical measure that compares the risk of an event or outcome in one group with the risk of the same event or outcome in another group. It is usually expressed as a ratio of the probability of an event occurring in the exposed group to the probability of the same event occurring in the unexposed group.**[7]**

2.2. ACADEMIC PAPERS

****Preface****

As there is a lack of academic research connecting debt to the occurrence of cancer directly, and none that utilize a similar approach of data/visual analytics to investigate the connection between debt and cancer incidence, I am suggesting an alternative approach. Instead, I will attempt to establish a cyclic relationship between debt and stress/depression/anxiety, and subsequently connect stress/depression/anxiety to cancer to establish the proposed relationship. This relationship will follow the notion that **"if A leads to B, and B leads to C, then we can conclude that A leads to C."**

The circular relationship between these factors is depicted in the illustration to the right:



****Acknowledgment****

The 3 academic papers below will serve as a validation for the three bubbles that are labeled 1-3 in the illustration mentioned above. However, as I have already covered these 3 papers in my progress report, I will instead discuss 2 further papers related to the data/visual analytics solution I am proposing.

Previously discussed papers:

1. The high price of debt: Household financial debt and its impact on mental and physical health [3]
2. Depression and anxiety in relation to cancer incidence and mortality: a systematic review and meta-analysis of cohort studies [7]
3. Cancer Diagnoses and Household Debt Overhang [8]

New academic papers:

1. Psychological stress and breast cancer incidence: A systemic review [2]

- The knowledge accumulated in the 20th century regarding stress and its mechanism of action combined with the enthusiasm regarding the impact of stress hormones in cancer development has led to numerous research studies**. Strong, but empirical hypotheses sustaining stress cancerogenesis have appeared.** (Chiriac et al., 2018, p. 18) **[2].**
- More than two-thirds of the articles analyzed stress produced by important life events, such as: the death of a husband, child, parent or close friend, divorce, **financial problems**, personal or familial medical problems and living in a political or cultural difficult period or region (Chiriac et al., 2018, p. 23) **[2].**
- The qualitative analysis has **shown a possible association between stressful events and breast cancer incidence**. (Chiriac et al., 2018, p. 25) **[2].**

2. Work stress and the risk of cancer: A meta-analysis of observational studies [1]

- Here, meta-analysis of data on 281,290 study participants and 9,090 incident cancer cases reveals significant associations **between work stress and increased risk of colorectal, esophageal, and lung cancers.** (Yang et al., 2018, p. 2391) **[1].**
- In conclusion, our meta-analysis shows that **work stress is an important risk factor for lung cancer, colorectal cancer, esophagus cancer, and overall cancer**. (Yang et al., 2018, p. 2399) **[1].**

3. WORKFLOW

- **STRATEGIES.**

There were several feasible approaches in terms of statistical modelling, each possessing unique advantages and disadvantages. Some of the possible techniques include developing **simple linear regression models**, utilizing **nonlinear regression** or **multiple regression models**, among others.

Simple Linear Regression

Simple linear regression is a statistical technique that models the relationship between a dependent variable and a single independent variable by fitting a linear equation. It assumes a linear relationship between the two variables, normal distribution of residuals, and constant variance. Its **advantages** include simplicity, interpretability, and efficiency, but its **disadvantages** include assumptions of linearity, sensitivity to outliers, and independence of observations.

Multiple Regression

Multiple linear regression is a technique used to model the relationship between a dependent variable and two or more independent variables. Its **advantages** include flexibility, statistical inference, prediction, and control of confounding variables, while its **disadvantages** include sensitivity to violations of the linearity assumption, overfitting, multicollinearity, and model selection.

Nonlinear Regression

Nonlinear regression is a technique used to model the relationship between a dependent variable and one or more independent variables when the relationship is not linear. Its **advantages** include flexibility, accurate predictions, and statistical inference, while its **disadvantages** include model selection, overfitting, and computational intensity.

- **IMPLEMENTATIONS.**

In my implementation, I opted to create both linear and multiple regression models. Since we are interested in investigating a path-based relationship between debt and mental health conditions in relation to cancer incidence, examining these independent variables both individually and together in a multiple regression model is essential if we are to establish a causal relationship.

We utilized a basic linear regression model with consumer credit and mortgage liabilities to disposable income or credit market debt to disposable income as the independent variable, and cancer incidence rate per 100,000 population in Canada as the dependent variable. Additionally, multiple linear regression models were tested with debt and mental health conditions (stress) in relation to cancer incidence.
Moreover, our cancer dataset enabled us to create subsets based on sex, age range, category of cancer, and geographical location (i.e., all of Canada or by province) over multiple years. Therefore, I conducted subset testing to determine whether any connections could be identified between the debt dataset and subsets of the cancer incidence dataset.

- **LIMITATIONS.**

I considered using non-linear regression but opted for simple linear regression as it requires fewer assumptions and less computation. It is also more intuitive to interpret the results of simple linear regression, which can be expressed in a simple equation. However, the choice between the two methods ultimately depends on the research question, the nature of the data, and the assumptions made about the relationship between the variables. If there is evidence of a non-linear relationship or a more complex pattern in the data, non-linear regression may be more suitable.

4. RESULTS

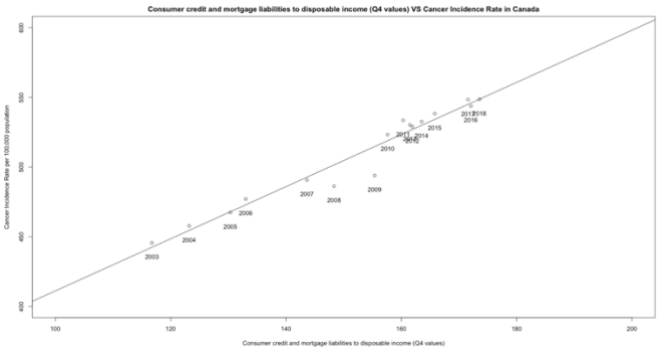
4.1 EXPERIMENTS.

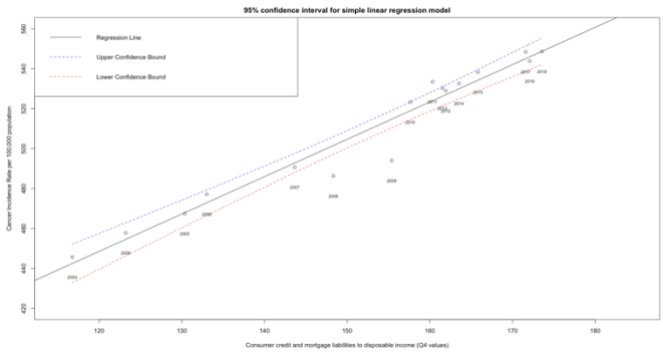
Since there were no experiments conducted to collect the data, I utilized the statistical method of linear and multiple regression models to analyze my data. In addition, I also compared scatterplots of individual cancer incidence and debt variables as the dependent variable over independent variable being year (2003-2018) after modifying the cancer incidence data set.

4.2. QUANTITATIVE.

The inspiration for this study came when I plotted a line graph for cancer incidence over years (2003-2018). I noticed a clear spike around the year 2008, with a 7.61% increase in cancer incidence from 2008 to 2010, and a 5.93% increase in cancer incidence from 2009 to 2010 **[Figure 2 & 3, Appendix].**

Next, I created a Linear Regression Model to examine the relationship between Consumer Credit and Mortgage Liabilities to Disposable Income (Q4 Values) and Cancer Incidence in Canada. A full list of summaries statistics can be found in **[Figure 7, Appendix]**. The **null hypothesis** was that there is **no significant relationship** between Consumer Credit and Mortgage Liabilities to Disposable Income and Cancer Incidence in Canada, while the **alternative hypothesis** was that there **is a significant relationship** between these variables. The **test-statistic** was **16.48**, and the **p-value of the t-statistic was 1.45e-10**, which is less than the **threshold value of 0.05**. Additionally, **the model P-value was 1.453e-10**, which is also **less than the threshold value of 0.05**. Based on these results, we can **reject the null hypothesis in favor of the alternative hypothesis**, which suggests that there is a significant relationship between Consumer Credit and Mortgage Liabilities to Disposable Income and Cancer Incidence in Canada. This plot is shown on the bottom left below:

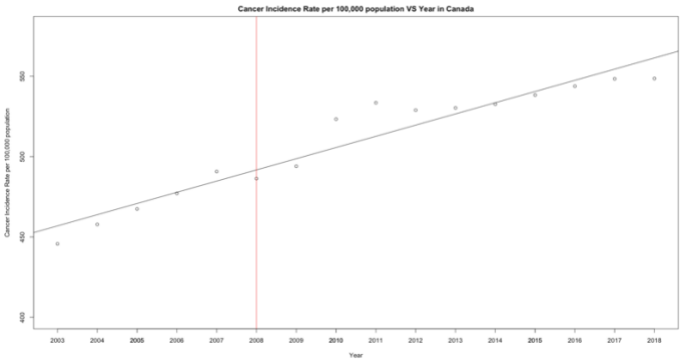


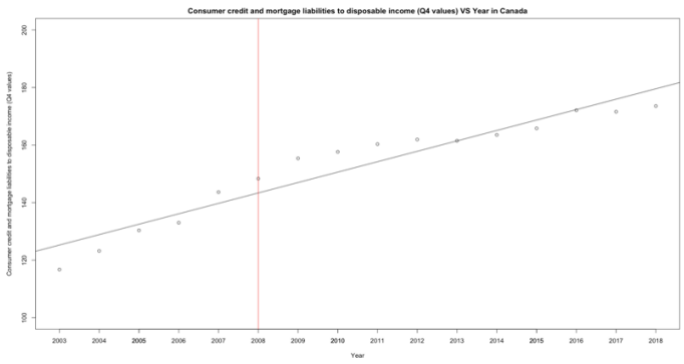


We also created a 95% Confidence Intervals for Linear Regression Model of Consumer Credit and Mortgage to Disposable Income (Q4 Values) VS Cancer Incidence in Canada **[Figure 8, Appendix].** Some notable statistics calculated include that the 95% confidence interval for the **Consumer Credit Debt coefficient ranges from 1.63 to 2.11**. The interval for the intercept suggests **that if the Consumer Credit Debt predictor variable is held constant**, the **average value of the Cancer Incidence Rate should be somewhere between 186.90 and 261.54**. This plot is shown above on the top right.

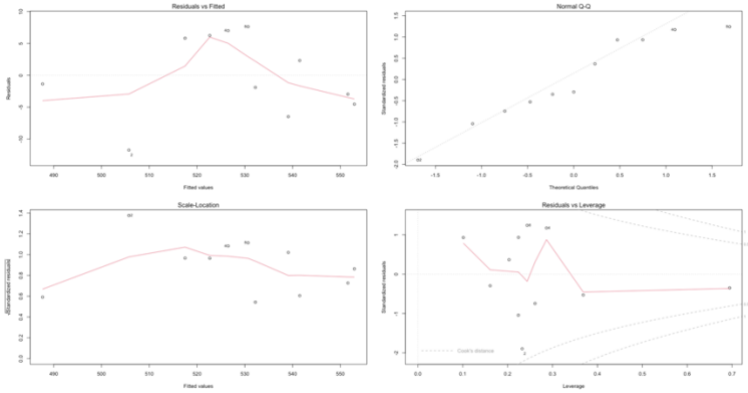
Similarly, calculations were performed for Linear Regression Model of Credit-Market Debt to Disposable Income (Q4 Values) VS Cancer Incidence in Canada, where we **also reject the null hypothesis** in favour of the alternative hypothesis that there is a significant relationship between Credit-Market Debt to Disposable Income and Cancer Incidence in Canada. The plot and summary statistics/interpretations can be found in **[Figure 9 & 10, Appendix].** Similarly, a 95% Confidence Intervals for Linear Regression Model of Credit-Market Debt to Disposable Income (Q4 Values) VS Cancer Incidence in Canada was calculated with summary statistics and interpretations found in **[Figure 11, Appendix].**

When examining plots of the three variables - consumer credit and mortgage liabilities to disposable income, credit market debt to disposable income, and cancer incidence (per 100,000 population) - with the year as the x-axis ranging from 2003 to 2018, it appears that all three plots follow a similar pattern with data points falling above and below the regression line in the same year(s). Notably, all three plots experienced a spike above the regression line around 2008, which coincided with the 2008 worldwide recession and subsequent Canadian recession, resulting in job losses, a housing market crash, a stock market crash, and an overall increase in debt across the country. My analysis suggests that the values were below the regression line around 2007 to 2009. This could indicate that the increase in debt was growing at a faster rate than cancer incidence during that period, which aligns with my hypothesis. Since a time-based dataset implies that debt must precede cancer if there is a causal relationship, this finding supports my reasoning. It takes time for cancer to develop, which explains why there was latency prior to the significant spike between 2009 and 2010. Plots for consumer credit and mortgage liabilities to disposable income and cancer incidence plotted over years are **shown below**, however full graph comparisons for the three variable scatterplots over years compared to their linear relation models can be found in **[Figure 14 & 15, Appendix].**





I aimed to develop a multiple regression model that encompasses both debt and mental health conditions and their connection to cancer incidence. This model incorporates two independent variables: mortgages, obtained from the national balance sheet, and the percentage of Canadians who have reported experiencing significant levels of stress, across several years within the context of Canada. Our dataset exhibits normal distribution, as evidenced by the Normal QQ plot depicted below on the left and in **[Figure 13, Appendix].**



Variable	Coefficient	Test-Statistic	P-value
Intercept	135.1	N/A	N/A
Mortgage Loans	0.0001121	-7.627	6.15e-05
Perceived Life Stress	11.47	-3.062	0.0155

Some notable statistics for the multiple regression model are shown above on the top right, as well as in **[Figure 16, Appendix].** The **null hypothesis** is that there is **no significant relationship** between Mortgage Loans (NBS Q4 Values) and Perceived Life Stress and Cancer Incidence in Canada, while the **alternative hypothesis** is that there is **a significant relationship** between these variables. The **test-statistic for Mortgage Loans is 7.627** and for **Perceived Life Stress is 3.062**. The **"p-value of the t-statistic" for Mortgage Loans is 6.15e-05**, which **is less than the threshold value of 0.05**, and for **Perceived Life Stress it is 0.0155**, also **less than the threshold**. Since both the p-values of the t-statistics are below the threshold, it means that **using mortgage loans and perceived life stress together is significantly better than using only one of the independent variables alone to predict cancer incidence**. Similarly, the **model P-value is 8.463e-05 = 0**, which is **less than the threshold value of 0.05**, meaning we can **reject the null hypothesis in favor of the alternative hypothesis**. That is, there is a significant relationship between Mortgage Loans (NBS Q4 Values) and Perceived Life Stress and Cancer Incidence in Canada.

4.3. INSIGHTS.

My results suggest statistically significant positive correlation between Consumer Credit and Mortgage Liabilities to Disposable Income, as well as Credit Market Debt to Disposable Income, and Cancer Incidence. Additionally, the study found a statistically significant positive correlation between Mortgage Loans and Perceived Life Stress in relation to Cancer Incidence. The plots for Debt and Cancer both followed a similar pattern, with data points falling above and below the regression line in the same year(s). There was a spike in the plots above the regression line around 2008, which coincides with the 2008 Canadian recession. This recession resulted in job losses, a housing market crash, a stock market crash, and an overall increase in debt across the country.

4.4. OUTCOMES.

There are several ways to measure the success of our data analysis study. Firstly, the study achieved most of the goals I had set out in section 1.2.2, Goals & Research Questions. Another measure of success is the quality of the analysis. The fact that both my linear and multiple regression models indicated that we could accept the alternative hypothesis, which was my primary objective, speaks to the accuracy and completeness of the data, the appropriateness of the statistical methods used, and the validity and reliability of the findings. Finally, we can measure success in terms of the impact of the findings on the research field. As my research topic is novel, the impact of my findings carries even more weight. Debt in relation to cancer is a lightly explored topic in the cancer research space, thus my study sits at the forefront of this topic's related work. Considering all these factors, I believe that my results were successful.

5. DISCUSSION

5.1. APPROACH.

The approach I adopted was promising. I created individual scatterplots to showcase linear relationships between each variable (Consumer Credit Debt, Credit-Market Debt, and Cancer Incidence). The standardized range of years (2003-2018) served as the independent variable across all plots, which helped to bridge the gap between my various datasets. This enabled me to identify interesting trends and establish a clear, easy-to-understand basis. I used this basis to justify the establishment of a simple linear relationship between debt and cancer incidence, which I achieved through a simple linear regression model. This multi-step approach aligned with how I collected and presented supporting academic papers, creating a mutually reinforcing structure. Additionally, using this sound linear regression model, I was able to extend my approach to multiple regression, where I found a statistically significant positive correlation between mortgage loans and perceived life stress in relation to cancer incidence.

Since my cancer dataset allowed me to create subsets based on sex, age range, cancer category, and geographical location (such as Canada as a whole or by province) across multiple years, I could have researched more to identify specific cancer types and relate them to debt to see if stronger evidence to support my hypothesis (debt causally related to cancer) existed. Similarly, I could have researched more to identify specific types of debt related to cancer incidence to see if stronger evidence to support my hypothesis existed. This could lead to exploring multiple independent variables, in addition to debt and stress, and their combined relationship with cancer incidence.

5.2. FUTURE WORK.

Some follow-up work that should be done next include:

Short-term:

- Explore more subsets of the each of my datasets. For example, explore subsets of the cancer dataset, such as different combinations of cancer type, age group, and sex, to identify any potentially interesting trends.
- Conducting further investigation into finding academic research that support the causality between debt and cancer.

Medium-term:

- Since my analysis only explored data up till 2018, extending the study to look at 2019 and onwards could provide further evidence of the relationship between debt and cancer, especially in light of the recent pandemic.
- In terms of multiple regression, future work could include finding or collecting data for different mental health conditions such as depression and anxiety and analysing its relationship in conjunction with debt in relation to cancer.

Long-term:

- Considering the speculation that Canada may face another recession in the near future (in 2023 or 2024), follow-up work can be conducted to extend the existing data on the relationship between debt, stress, and cancer incidence, and to predict values of cancer incidence based on increasing debt and stress variable values. This may potentially provide speculative values of cancer incidence rates that could occur if severe country-wide spikes in debt are experienced. One goal of this objective would be to compare the predicted values with the actual values of debt, stress, and cancer incidence following the completion and release of 2023 and 2024 debt/stress/cancer statistics, which may take a few years before being publicly released.

6. CONCLUSION

I expect to utilize a combination of data science skills such as data collection, cleaning, visualization, and analysis to complete my project. Through my research topic, I anticipate gaining insight into the significance of relationships beyond a one-to-one basis. It is important to note that a lack of relationship between two variables does not necessarily imply that they are not related. Rather, it may be necessary to establish a pathway from the independent variable (cause) to the dependent variable (effect) to uncover meaningful relationships. By tracing the root cause (e.g., debt) and following its path (e.g., from debt to stress to a weaker immune system to increased susceptibility to cancer), more meaningful relationships can be established. We can conclude that debt and cancer may be causally related, with supporting evidence in the academic papers featured in this study as well as my statistical analysis.

7. REFERENCES

[1] Yang, T., Qiao, Y., Xiang, S., Li, W., Gan, Y., & Chen, Y. Work stress and the risk of cancer: A meta-analysis of observational studies. *144*(10), 2390–2400, (2018), *International Journal of Cancer*. https://doi.org/10.1002/ijc.31955

[2] Chiriac, V.-F., Baban, A., & Dumitrascu, D. L. Psychological stress and breast cancer incidence: A systematic review. *91*(1), 18–26. (2018), Medicine *and Pharmacy Reports*. https://doi.org/10.15386/cjmed-924

[3] E., Nandi, A., Adam, E. K., & McDade, T. W. The high price of debt: Household financial debt and its impact on mental and Physical Health. *91*, 94–100, (2013), *Social Science & Medicine*. https://doi.org/10.1016/j.socscimed.2013.05.009

[4] Amit, N., Ismail, R., Zumrah, A. R., Mohd Nizah, M. A., Tengku Muda, T. E., Tat Meng, E. C., Ibrahim, N., & Che Din, N. Relationship between debt and depression, anxiety, stress, or suicide ideation in Asia: A systematic review. *11*, (2020), *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2020.01336

[5] Sebastien LaRochelle-Cote & Jason Gilmore, *Canada's employment downturn,* 1-8, December/2009, Statistics Canada. https://www150.statcan.gc.ca/n1/en/pub/75-001-x/2009112/pdf/11048-eng.pdf?st=IrOErz1X

[6] Philip Cross, *Canadian Economic Observer - 2008 in Review,* 1-22, 2009, Statistics Canada https://www150.statcan.gc.ca/n1/en/pub/11-010-x/2009004/article/10848-eng.pdf?st=7k91ORCM

[7] Yun-He Wang, Jin-Qiao Li, Ju-Fang Shi, Jian-Yu Que, Jia-Jia Liu, Julia M. Lappin, Janni Leung, Arun V. Ravindran, Wan-Qing Chen, You-Lin Qiao, Jie Shi, Lin Lu & Yan-Ping Bao. Depression and anxiety in relation to cancer incidence and mortality: a systematic review and meta-analysis of cohort studies, 25:1487–1499, (2020) *Molecular Psychiatry*. https://pubmed.ncbi.nlm.nih.gov/31745237/

[8] Arpit Gupta, Edward R. Morrison, *Cancer Diagnosis and Household Debt Overhang,* 1-36, May/2015, Fred Hutchinson Cancer Research Center. https://static1.squarespace.com/static/56086d00e4b0fb7874bc2d42/t/56088819e4b062e5cb9a2eb7/1443399705239/CancerLeveragePaper.pdf

[9] Roxanne Garaszczuk, Jean H. E. Yong, Zhuolu Sun & Claire de Oliveira, *The Economic Burden of Cancer from a Societal Perspective,* 2735-2748, April 2022, Current Oncology. https://www.mdpi.com/1718-7729/29/4/223

[10] Canadian Cancer Society. Canadian Cancer Statistics 2021. Toronto, ON, 2021, Canadian Cancer Society. https://cdn.cancer.ca/-/media/files/research/cancer-statistics/2021-statistics/2021-pdf-en-final.pdf

[11] John Elflein, Percentage of Canadians who perceived quite a lot of life stress from 2003 to 2021, Sept/2022, Statista. https://www.statista.com/statistics/434081/share-of-canadians-perceiving-their-life-stress-as-quite-a-lot/

8. APPENDIX

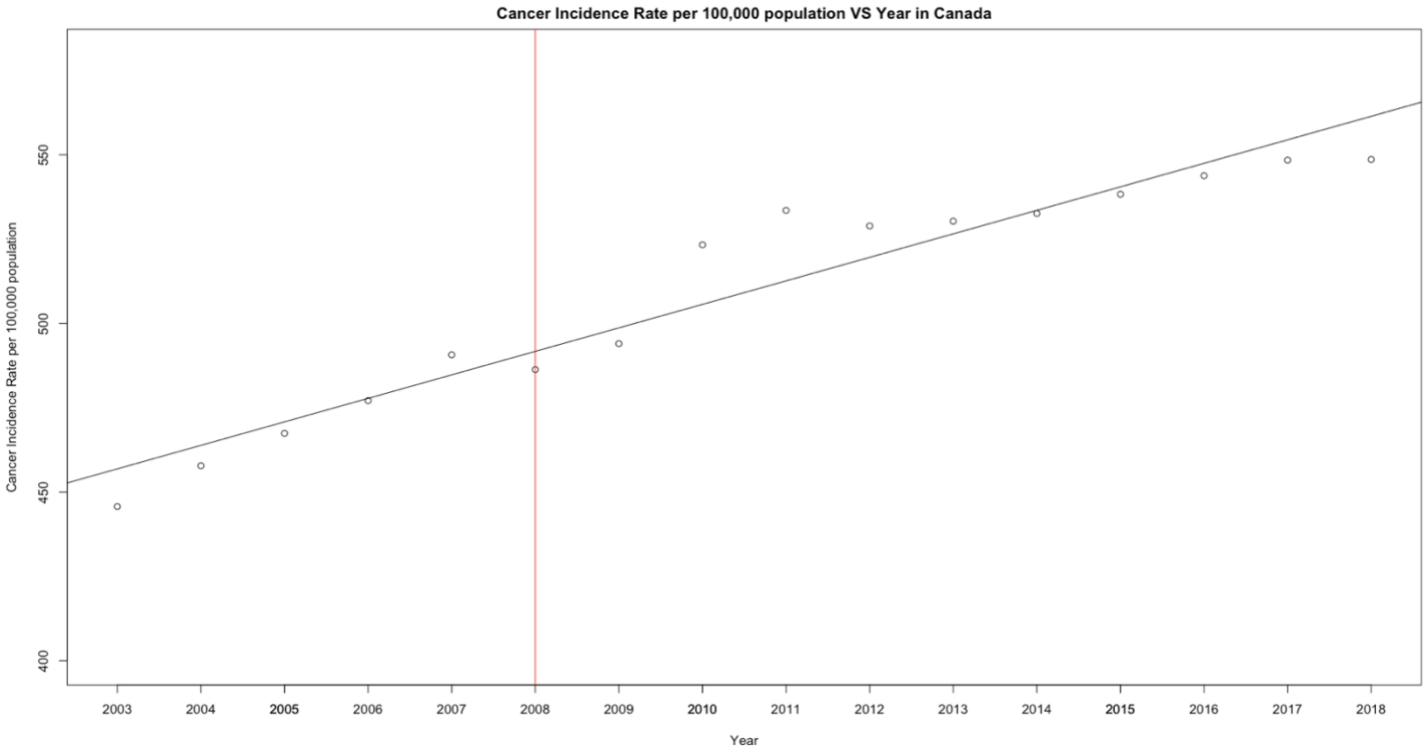


FIGURE 1: Cancer Incidence Rate Per 100000 Population VS Year in Canada

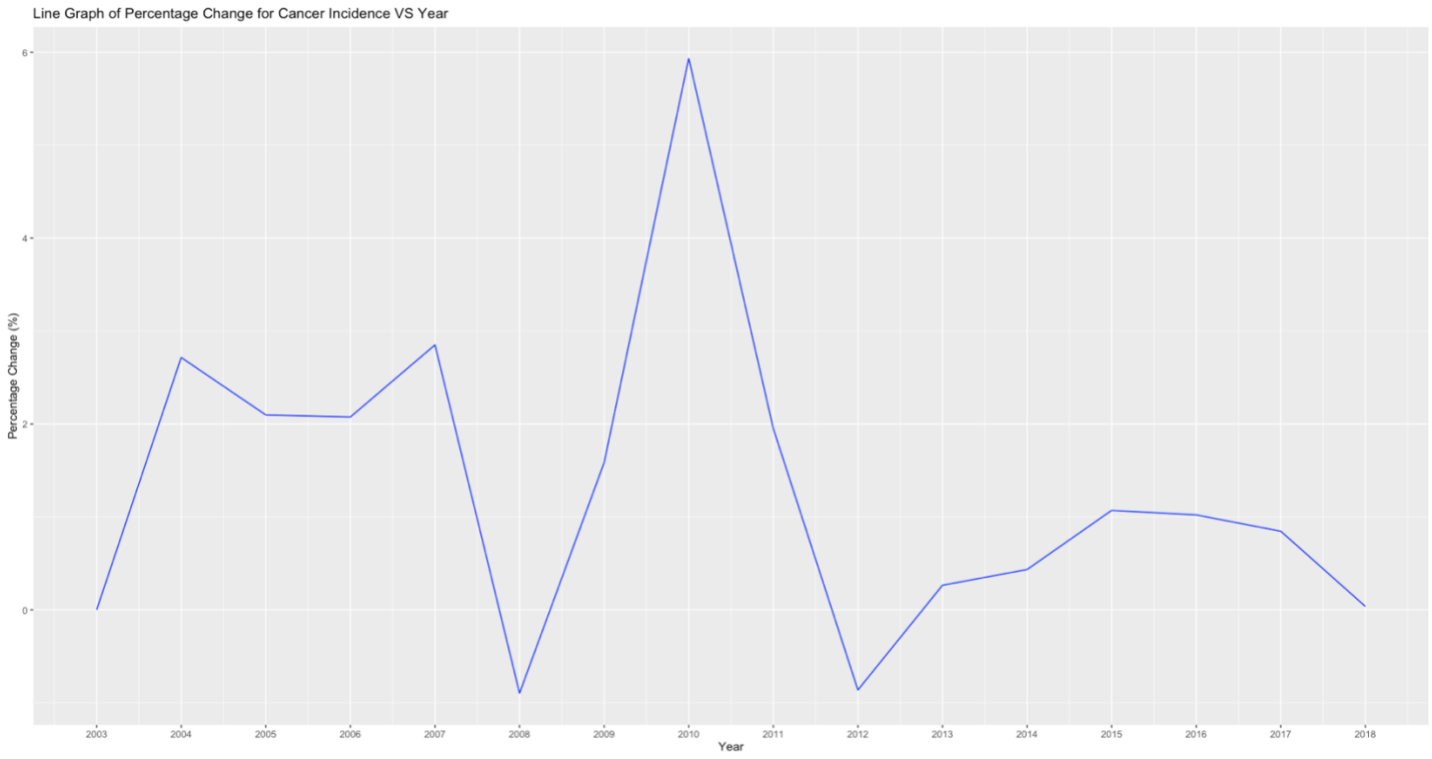


FIGURE 2: Line Graph of Percentage Change for Cancer Incidence VS Year in Canada

Calculation	Formula/Values	Interpretation
Percentage Increase from 2008 to 2010	Difference: 37, Initial Value: 486.3	The value increased by 7.61% from 2008 (486.3) to 2010 (523.3).
	Difference/Initial Value: 0.076084721365412	
	Difference/Initial Value * 100: 7.61%	
Percentage Increase from 2009 to 2010	Difference: 29.3, Initial Value: 494	The value increased by 5.93% from 2009 (494) to 2010 (523.3).
	Difference/Initial Value: 0.059311740890688	
	Difference/Initial Value * 100: 5.93%	

FIGURE 3: Summary Statistics and Interpretations for Line Graph of Percentage Change for Cancer Incidence VS Year in Canada

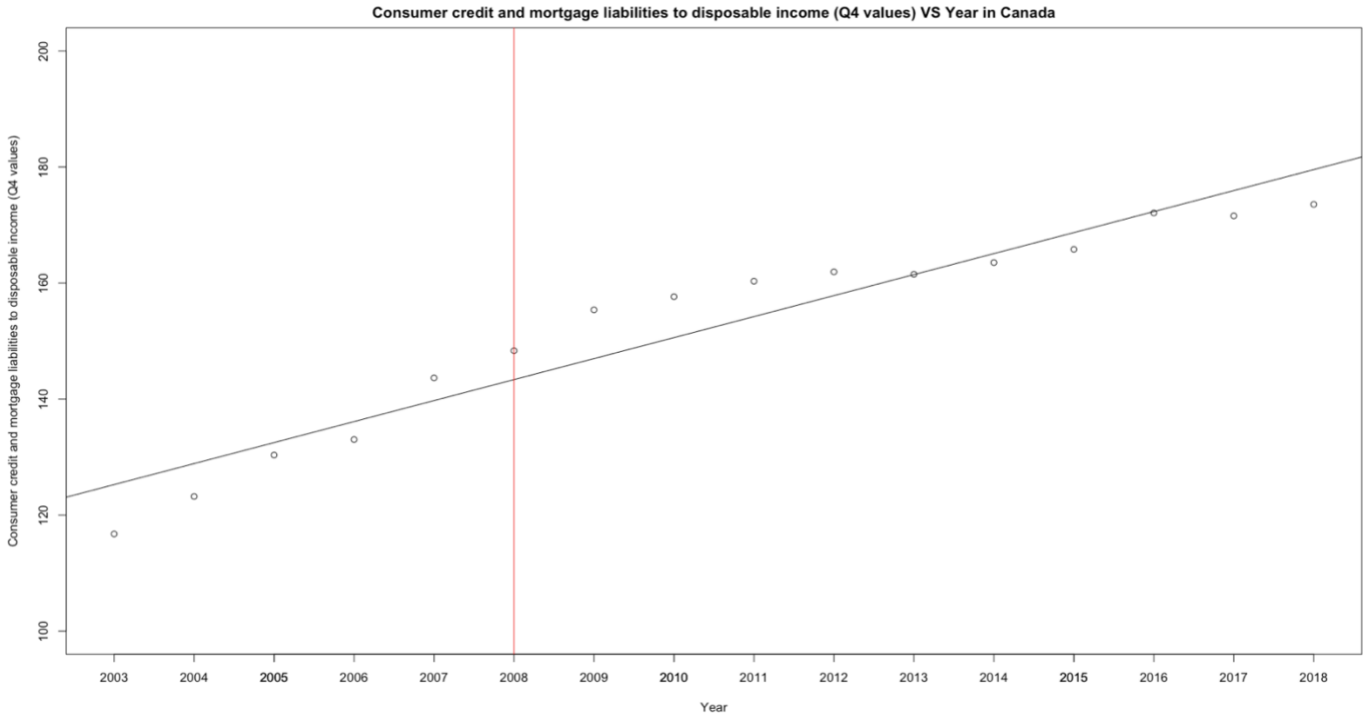


FIGURE 4: Consumer Credit and Mortgage to Disposable Income (Q4 Values) VS Year in Canada

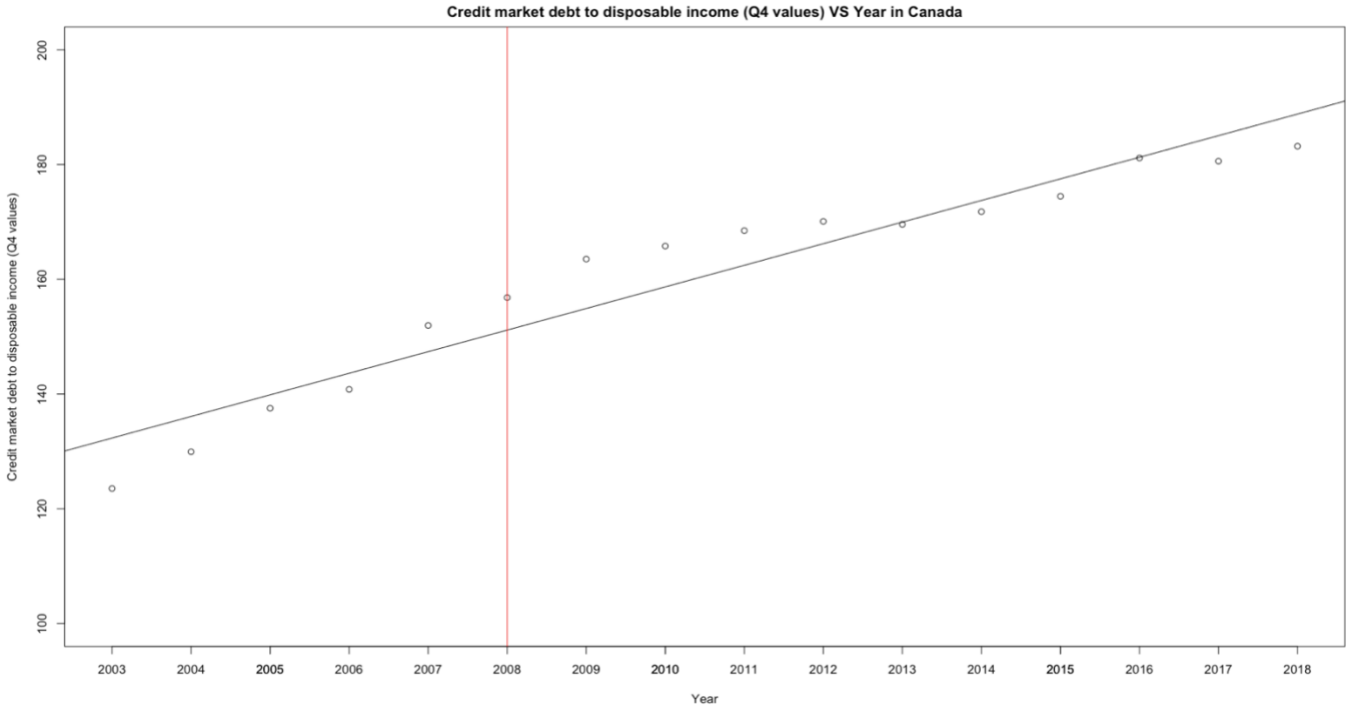


FIGURE 5: Credit-Market Debt to Disposable Income (Q4 Values) VS Year in Canada

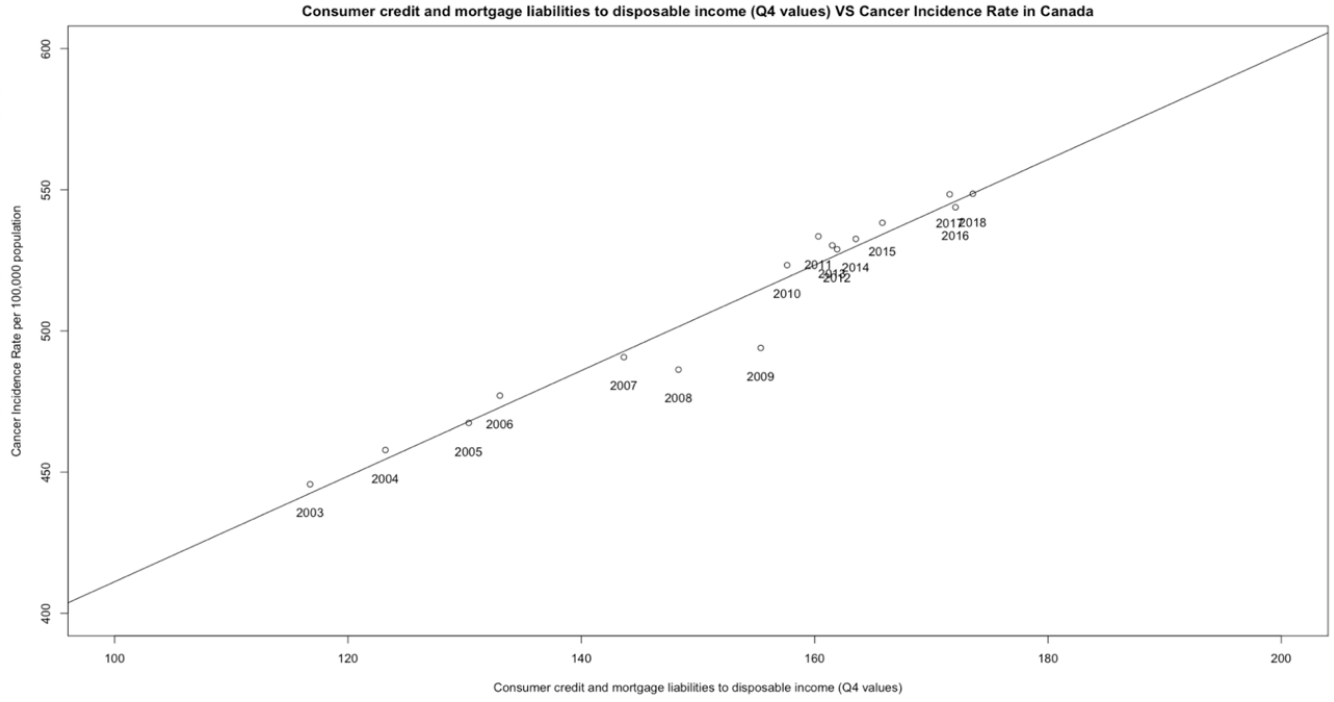


FIGURE 6: Linear Regression Model of Consumer Credit and Mortgage to Disposable Income (Q4 Values) VS Cancer Incidence in Canada

Quantitative Results	Formula/Values	Interpretation
Regression Formula	y = 224.22 + (1.87 * x)	The model predicts that Cancer Incidence = 224.22 + (1.87 * Consumer Credit Ratio)
Residuals	Min: -20.711, 1Q: -0.907, Median: 2.931, 3Q: 4.099, Max: 9.554	The median is close to 0, indicating that the model is not skewed in any particular direction.
Slope	1.87	A 1 unit increase in Consumer Credit Debt Ratio is associated with a 1.87 increase in Cancer Incidence.
Std. Error (intercept)	17.4002	Roughly 95% of the observations should fall within +/- two standard errors of the regression.
Std. Error (weight)	0.1134	
t-value (intercept)	12.89	The intercept is statistically significant with a t-value of 12.89.
t-value (weight)	16.48	The weight is statistically significant with a t-value of 16.48.
p-value	1.453e-10 = 0 < 0.05(threshold)	The p-value of 1.453e-10 is less than the threshold of 0.05, indicating that the model is statistically significant.
Residual Standard Error	7.905 on 14 degrees of freedom	The residual standard error is 7.905 on 14 degrees of freedom.
Adjusted R-Squared	0.9475	94.75% of the variability observed in the target variable (Cancer Incidence) is explainable by the regression model.
F-Statistic	271.7 on 1 and 14 DF	The F-statistic of 271.7 on 1 and 14 degrees of freedom indicates that the regression model is statistically significant.

FIGURE 7: Summary Statistics and Interpretations for Linear Regression Model of Consumer Credit and Mortgage to Disposable Income (Q4 Values) VS Cancer Incidence in Canada

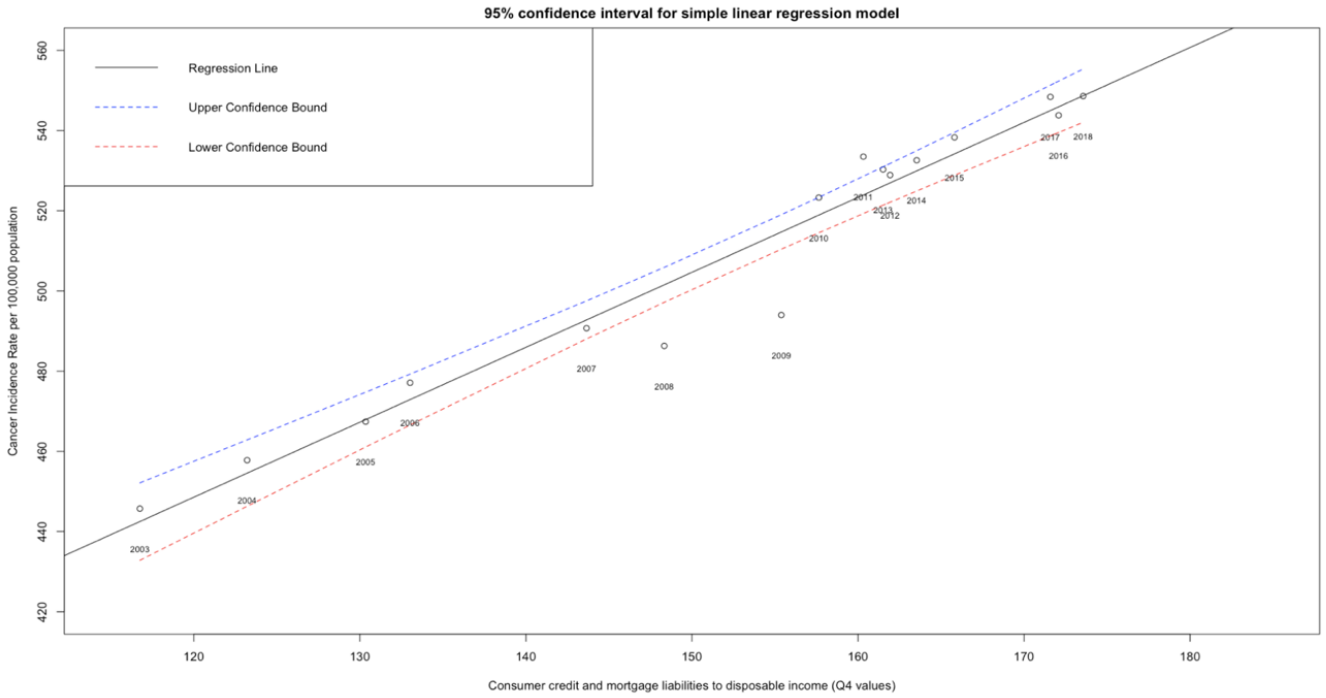


FIGURE 8: 95% Confidence Intervals for Linear Regression Model of Consumer Credit and Mortgage to Disposable Income (Q4 Values) VS Cancer Incidence in Canada. The 95% confidence interval for the Consumer Credit Debt coefficient ranges from 1.63 to 2.11. The interval for the intercept suggests that if the Consumer Credit Debt predictor variable is held constant, the average value of the Cancer Incidence Rate should be somewhere between 186.90 and 261.54.

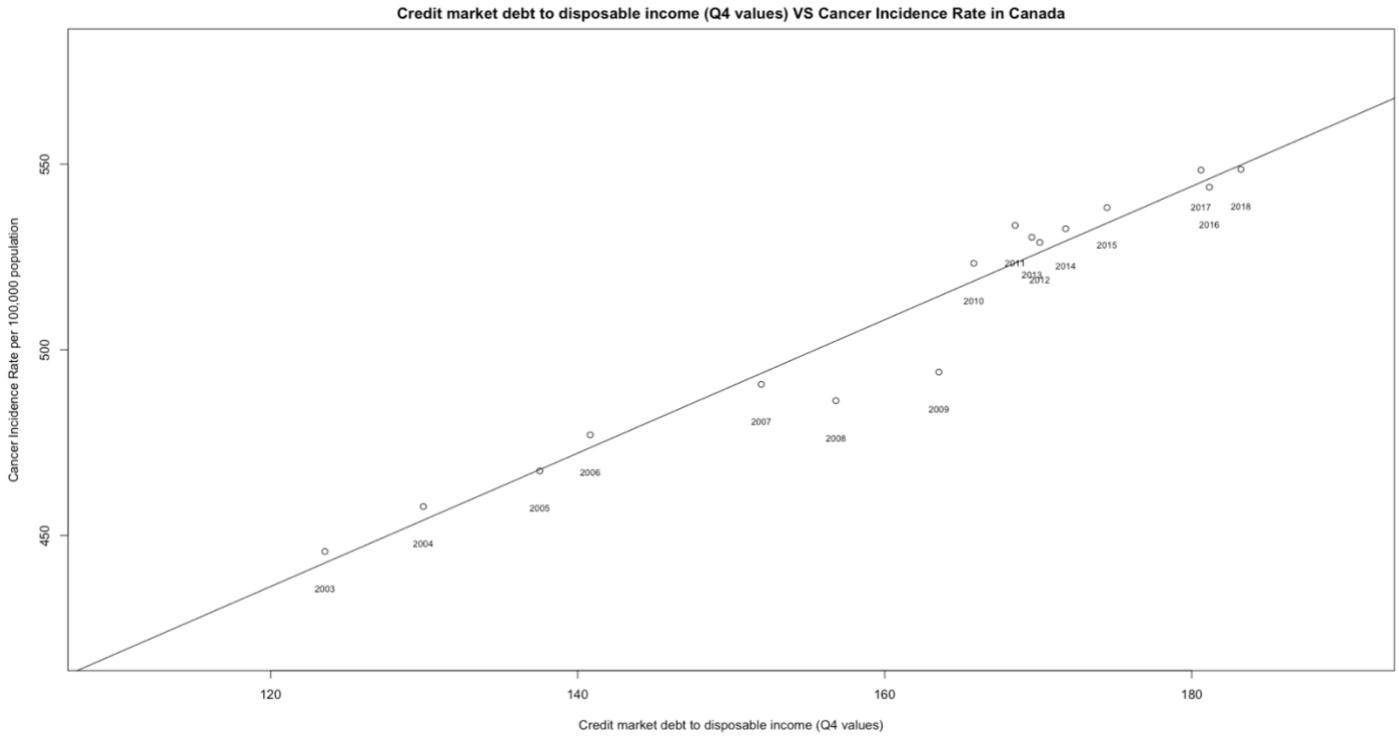


FIGURE 9: Linear Regression Model of Credit-Market Debt to Disposable Income (Q4 Values) VS Cancer Incidence in Canada

Quantitative Results	Formula/Values	Interpretation
Regression Formula	y = 220.695 + (1.796 * x)	The model predicts that Cancer Incidence = 220.695 + (1.796 * Credit-Market Debt Ratio)
Residuals	Min: -20.463, 1Q: -1.477, Median: 3.174, 3Q: 3.801, Max: 10.127	The median is close to 0, implying that the model is not skewed in one direction or another.
Slope	1.796	A 1 unit increase in Credit-Market Debt Ratio results in a 1.796 increase in Cancer Incidence.
Std. Error (intercept)	18.1325	Roughly 95% of the observations should fall within +/- two standard errors of the regression, which is a quick approximation of a 95% prediction interval.
Std. Error (weight)	0.1122	
t-value (intercept)	12.17	The intercept is statistically significant with a t-value of 12.17.
t-value (weight)	16.01	The weight is statistically significant with a t-value of 16.01.
p-value	2.143e-10 = 0 < 0.05(threshold)	The p-value of 2.143e-10 is less than the threshold of 0.05, indicating that the model is statistically significant.
Residual Standard Error	8.127 on 14 degrees of freedom	The residual standard error is 8.127 on 14 degrees of freedom.
Adjusted R-Squared	0.9445	94.45% of the variability observed in the target variable (Cancer Incidence) is explainable by the regression model.
F-Statistic	256.3 on 1 and 14 DF	The F-statistic of 256.3 on 1 and 14 degrees of freedom indicates that the regression model is statistically significant.

FIGURE 10: Summary Statistics and Interpretations for Linear Regression Model of Credit-Market Debt to Disposable Income (Q4 Values) VS Cancer Incidence in Canada

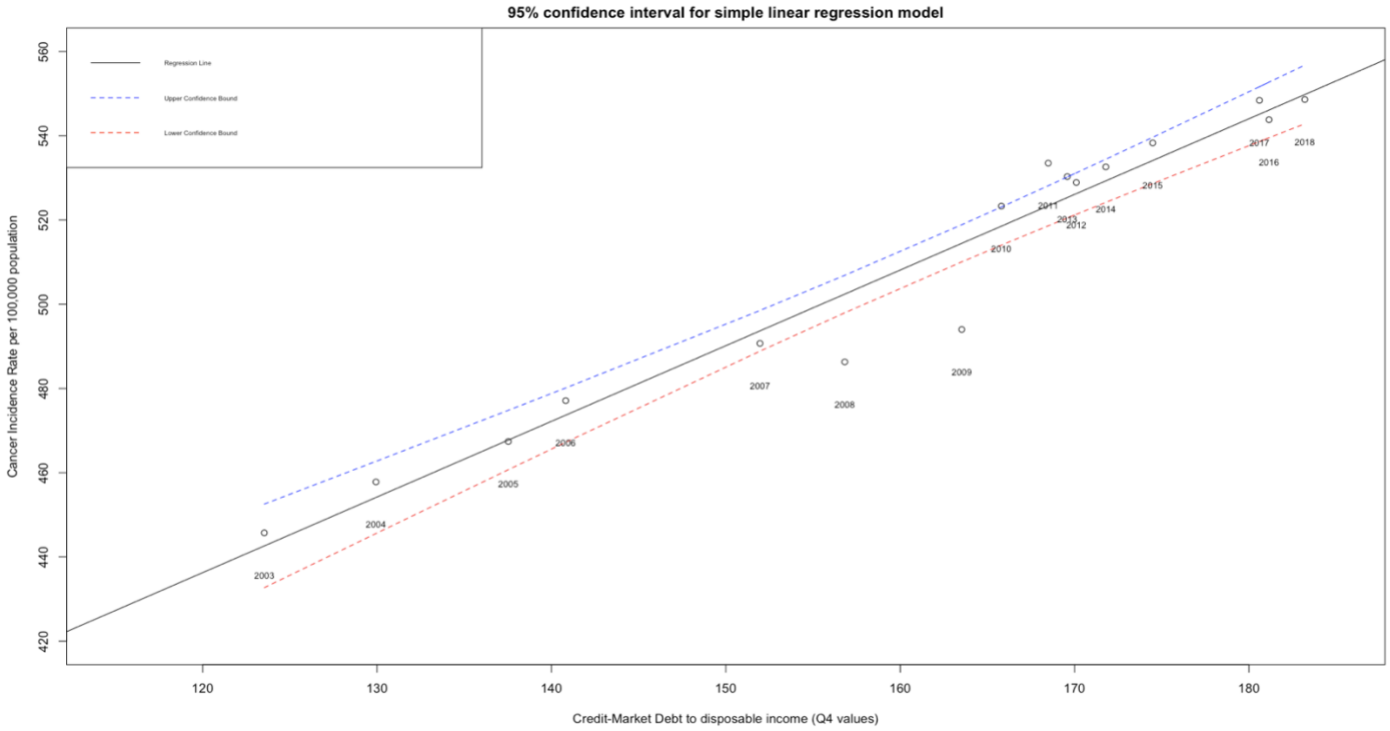


FIGURE 11: 95% Confidence Intervals for Linear Regression Model of Credit-Market Debt to Disposable Income (Q4 Values) VS Cancer Incidence in Canada. The 95% confidence interval for the Credit-Market Debt coefficient ranges from 1.56 to 2.04. The interval for the intercept suggests that if the Credit-Market Debt predictor variable is held constant, the average value of the Cancer Incidence Rate should be somewhere between 181.80 and 259.58.

FIGURE 12: Canadian Cancer Society - Canadian Cancer Statistics Report 2021 Summary Infographic [11]

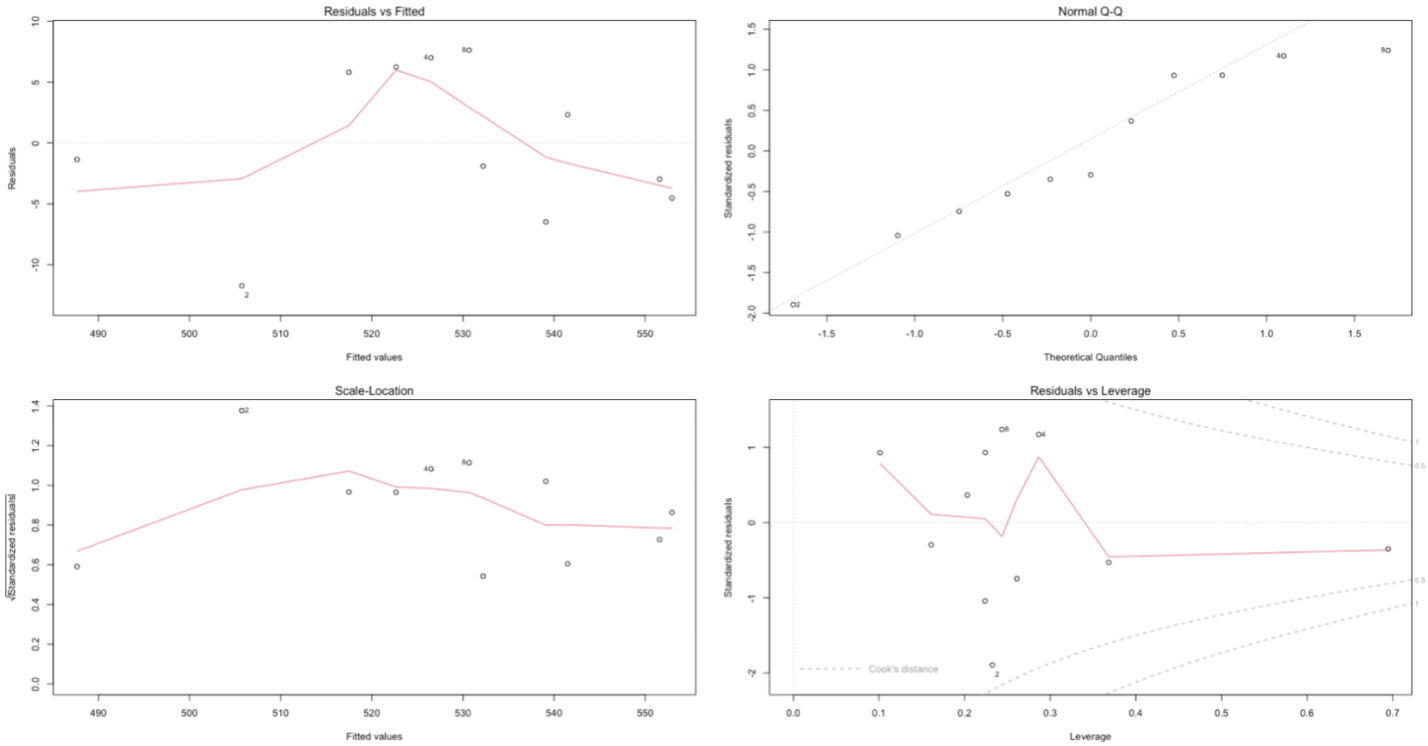


FIGURE 13: Multiple Regression Model for Mortgage Loans (NBS Q4 Values) and Perceived Life Stress VS Cancer Incidence

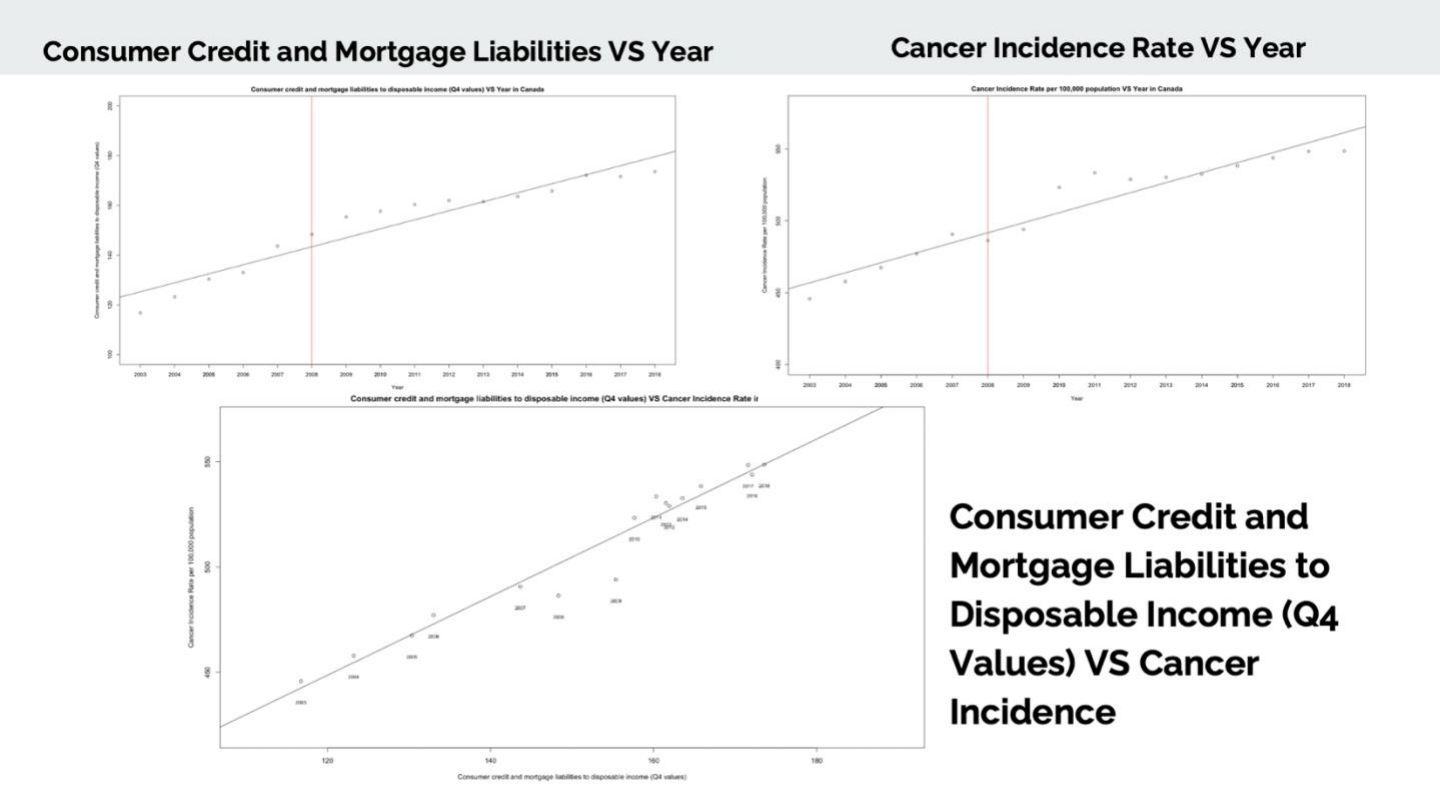


FIGURE 14: Graph Comparison of Individual variable scatterplots vs linear regression model of both variables. Top left graph shows the plot of Cancer Incidence per 100000 population VS Year (2003-2018) in Canada Top right graph shows the plot of Consumer Credit and Mortgage Liabilities to Disposable Income (Q4 Values) VS Year (2003-2018) in Canada. Bottom graph shows the linear regression model between Consumer Credit and

Mortgage Liabilities to Disposable Income (Q4 Values) VS Cancer Incidence per 100000 population with data from 2003-2018 in Canada. As you can see that both top graphs show very similar trends, and this trend transfers into the bottom model.

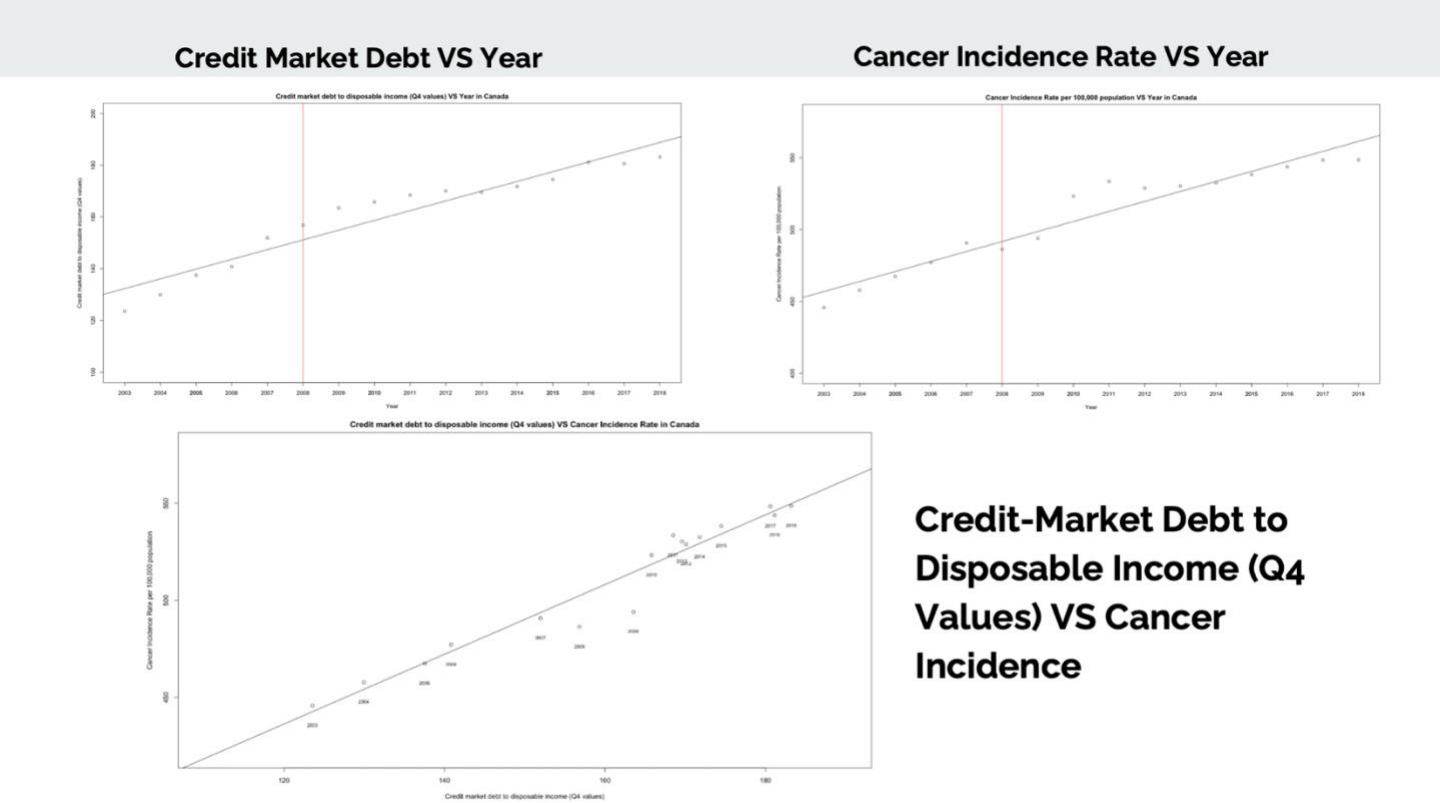


FIGURE 15: Graph Comparison of Individual variable scatterplots vs linear regression model of both variables. Top left graph shows the plot of Cancer Incidence per 100000 population VS Year (2003-2018) in Canada Top right graph shows the plot Credit-Market Debt to Disposable Income (Q4 Values) VS Year (2003-2018) in Canada. Bottom graph shows the linear regression model between Credit-Market Debt to Disposable Income (Q4 Values) VS Cancer Incidence per 100000 population with data from 2003-2018 in Canada.
As you can see that both top graphs show very similar trends, and this trend transfers into the bottom model.

Variable	Coefficient	Test-Statistic	P-value
Intercept	135.1	N/A	N/A
Mortgage Loans	0.0001121	-7.627	6.15e-05
Perceived Life Stress	11.47	-3.062	0.0155

FIGURE 16: Table of summary statistics for the Multiple Regression Model for Mortgage Loans (NBS Q4 Values) and Perceived Life Stress VS Cancer Incidence

FOOTNOTES:

- DATA FOR CANCER INCIDENCE IS OBTAINED FROM ALL OF CANADA EXLUDING QUEBEC
- I ALSO USED Q4 DATA FOR DEBT VARIABLES AS I BELIEVE THAT Q4 WOULD BEST SYMBOLIZE THE AGGREGATE VALUE OF THE YEAR IT REPRESENTS.