Analyzing Effects of Cigarette Tax Cuts on Smoking Habits

Dhanraj Patel

2021-08-23

Contents

Abstract

Within this report we will investigate how the 1994 decrease of cigarette taxes in select provinces effected peoples smoking habits and analyze how their smoking patterns fluctuated within a year. We will use a survey conducted by Statistics Canada on a diverse group of smokers who were surveyed every three months for a year starting after the tax decrease. We used this data on a frequentist Poisson model and Bayesian (prior: Gamma, likelihood: Poisson) model to estimate the average number of cigarettes smoked by the survey participants in a day and week. Also, by using a Linear regression model we examined the how the total number of cigarettes smoked by all the participants varied throughout the survey. All three models showed similar results with an increase in the daily/weekly averages and total number of cigarettes smoked in the first three months of the survey; however, resulted in a decrease in the usage of cigarette in the last nine months. Despite the decrease in the smoking prevalence seen within the survey, the rate at which the smoking prevalence decreased did not reach the rate of provinces who did not lower the tax rate. From this, we conclude that decreasing the taxes on cigarettes slows down the decline in smoking rates by incentivizing people to smoke cigarettes by making them more affordable and accessible.

Introduction

Overview

All over the world, including Canada, smoking cigarettes continues to be a major problem for people of all ages. Due to the fact habitual smoking of cigarettes will eventually lead to the development of numerous different types of cancers there have been many information campaigns throughout the globe to inform people about the dangers of smoking. Aside from cancer risks, there are many other factors that influence an individuals smoking habits, with financial costs being one of the greatest factors. To decrease smoking rates, many countries include high rates of taxes on the sale of cigarettes to discourage smoking. In 1994 the select provinces lowered the taxes on cigarettes and in response Stats Canada on behalf of Health Canada conducted a year long survey on Canadians who identify as smokers where they were asked the same set of questions about their smoking habits every 3 months for a year. It is very important to note that the survey began closely after the tax cut was implemented in 1994.

The main purposes of the study was to measure:

- The amount of cigarettes smoked in the year
- Any changes the amount of cigarettes smoked in the year
- The effect that price has on the total amount of cigarettes smoked

The data and its subsequent analysis is important and relevant to not just Canada, but to the world as a whole as it helps us better understand the behaviors of smokers and the effects of factors such as price can have on smoking rates. By understanding these factors we can make more informed decisions on policies relating to smoking to decrease the amount of smokers worldwide.

Unerderstanding Terminiology

An important term to understanding about the survey conducted and the data is the idea of cycles of a survey. Within the survey Stats Canada in the year of 1994 surveyed the participants every three months with the same questions. Each round of surveys is knows as one cycle, with the entire survey consisting of 4 cycles in total. The purpose of the cycles is to record the smoking habits of the participants every three months to investigate if the decrease in cigarette taxes changed their smoking patterns.

Research Questions

In the survey a group of smokers were asked the same set of questions about the their smoking habits every three months over the course of the year. With the government having decreased the overall taxes on the sale of cigarettes, this report will investigate the overarching question on how the decrease of prices of cigarettes impacts the smoking behaviors and habits of Canadians?

To answer this overarching question we will be using different types of models to investigate different aspects of the data which will help us understand how the smoking habits of the survey participants have changed throughout the year. Using linear linear regression modeling we will investigate the question on how the total number of cigarette smoked in each cycle has changed throughout the year. In addition, using frequentist modeling we will analyze how the average amount of cigarettes smoked in one day has changed between the first and forth cycle. In this model we will be using hypothesis testing to see if there was no change between the first and forth cycle (null hypothesis) or if there was any significant change between the first and forth cycles (alternative hypothesis). Finally using Bayesian modeling we will examine the data to investigate how the average number of cigarettes smoked in a week changed between the first and forth cycle. We will be using a hypothesis testing to see if there was no significant change (null hypothesis) between the first and fourth cycles, or if there was any significant change between the cycles (alternative hypothesis). Both of the hypothesis tests will be 2 sided tests. By investigating these three questions with the different models we can understand the central question of the decrease on how the 1994 decrease on taxes on cigarettes effected the smoking habits of the participants of the survey.

Understanding the data

Data Collection Process

Before delving deeper into the methodologies and using the data to better understand the impact on the price of cigarettes and how they influences the behavior of smokers, it is first important to understand where the data came from. In 1994 select provinces in Canada lowered the tax rates on cigarettes and to study the impact it had on smokers Health Canada asked Stats Canada to conduct a survey due to the fact that Stats Canada is a government agency responsible for producing statistics to better understand Canada and its population. Stats Canada then conducted a year long study on a sample population of Canadian smokers. Stats Canada being a public agency then released the findings of the survey for the public to view. The data is available on http://odesi2.scholarsportal.info/webview/. The data used in this report can be found on that website (see citation(4)).

Cleaning of data

After retrieving the dataset it is important to clean the data so we are able to use it in models to investigate the effects that lowering the tax on cigarette has on smokers. In order to clean the data it was first important to eliminate the columns of data that would not be used for the purposes of our models. We want to count the amount of cigarettes smoked in all days of the week for all 4 of our cycles and to keep track of the survey ID number, sex, income level and age group of each participant. Thus, unrelated columns of data were removed.

Furthermore, to clean the data it was important to eliminate rows of data that contain missing information.In the survey it was possible for those who were surveyed to skip, state they did not know or refuse to answer the questions. So all rows that included missing answers were removed.

Finally for the data it was important to remove duplicate entries for those who participated in the survey. Using their unique ID number all duplicate entries were removed.

Important Variables

These are the important variables in the dataset that will be used in the graphical and numerical summaries as well as the models.

- SEX - The sex of the participant:
  - 1: Male
  - 2: Female
- AGEGP4 - The age interval of the participant:
  - 1: 15-19 years old
  - 2: 20-24 years old
  - 3: 25-64 years old
  - 4: 65+ years old
- C1INCAD - The income interval of the participant:
  - 1: Lower class
  - 2: Lower-middle class
  - 3: Upper-middle class
  - 4: Upper class

Table 1: Important variables for the first cycle of answers

Variable_name	Description
C1SMON	Number of cigarettes smoked on Monday during the first cycle of the survey.
C1STUE	Number of cigarettes smoked on Tuesday during the first cycle of the survey.
C1SWED	Number of cigarettes smoked on Wednesday during the first cycle of the survey.
C1STHU	Number of cigarettes smoked on Thursday during the first cycle of the survey.
C1SFRI	Number of cigarettes smoked on Friday during the first cycle of the survey.
C1SSAT	Number of cigarettes smoked on Saturday during the first cycle of the survey.
C1SSUN	Number of cigarettes smoked on Sunday during the first cycle of the survey.

From (Table_1) we can see the important variables for the first cycle of questions which contain the total number of cigarettes smoked in each day of the week.

Table 2: Important variables for the second cycle of answers

Variable_name	Description
C2SMON	Number of cigarettes smoked on Monday during the second cycle of the survey.
C2STUE	Number of cigarettes smoked on Tuesday during the second cycle of the survey.
C2SWED	Number of cigarettes smoked on Wednesday during the second cycle of the survey.
C2STHU	Number of cigarettes smoked on Thursday during the second cycle of the survey.
C2SFRI	Number of cigarettes smoked on Friday during the second cycle of the survey.
C2SSAT	Number of cigarettes smoked on Saturday during the second cycle of the survey.
C2SSUN	Number of cigarettes smoked on Sunday during the second cycle of the survey.

From (Table_2) we can see the important variables for the second cycle of questions which contain the total number of cigarettes smoked in each day of the week.

Table 3: Important variables for the third cycle of answers

Variable_name	Description
C3SMON	Number of cigarettes smoked on Monday during the third cycle of the survey.
C3STUE	Number of cigarettes smoked on Tuesday during the third cycle of the survey.
C3SWED	Number of cigarettes smoked on Wednesday during the third cycle of the survey.
C3STHU	Number of cigarettes smoked on Thursday during the third cycle of the survey.
C3SFRI	Number of cigarettes smoked on Friday during the third cycle of the survey.
C3SSAT	Number of cigarettes smoked on Saturday during the third cycle of the survey.
C3SSUN	Number of cigarettes smoked on Sunday during the third cycle of the survey.

From (Table_3) we can see the important variables for the third cycle of questions which contain the total number of cigarettes smoked in each day of the week.

Table 4: Important variables for the forth cycle of answers

Variable_name	Description
C4SMON	Number of cigarettes smoked on Monday during the forth cycle of the survey.
C4STUE	Number of cigarettes smoked on Tuesday during the forth cycle of the survey.
C4SWED	Number of cigarettes smoked on Wednesday during the forth cycle of the survey.
C4STHU	Number of cigarettes smoked on Thursday during the forth cycle of the survey.
C4SFRI	Number of cigarettes smoked on Friday during the forth cycle of the survey.
C4SSAT	Number of cigarettes smoked on Saturday during the forth cycle of the survey.
C4SSUN	Number of cigarettes smoked on Sunday during the forth cycle of the survey.

From (Table_4) we can see the important variables for the forth cycle of questions which contain the total number of cigarettes smoked in each day of the week.

Numerical summaries

After cleaning the dataset we are able to use numerical summaries to better understand the values and distribution of our data. It is important to focus on both the location and spread of the data

Comparing total number of cigarettes smoked in week for all cycles

The following table (labeled Table_5) displays numerical summaries of the total number of cigarettes smoked in a week for each cycle.

Table 5: The trimmed mean was trimmed by 10 percent

Cycle_number	Min	q1	Med	q3	Max	IQR	Mean	trimMean	var	sd
1	1	70	108	175	645	105	117.83	114.71	4271.78	65.36
2	1	70	106	175	525	105	118.40	116.35	3951.30	62.86
3	2	70	105	175	425	105	115.72	113.70	3805.98	61.69
4	1	70	105	165	560	105	111.84	108.99	4130.46	64.27

(Table_5) contains the numerical summaries of the total amount of cigarettes smoked in all of the cycles. From looking at the mean and trimmed means of all the cycles, we see that the total number of cigarettes smoked peaks at cycle two and then continues to decrease for cycles three and four. In addition, we see that compared to the first cycle, the second and third cycles have lower levels of variance compared with cycle one. Furthermore, the we notice how after decreasing, in the forth cycle the standard deviation and variace levels start increasing. Finally from the quartiles values we see the all the values from minimum to the third quartile remain close together for the first the cycles with a decrease in the values in the forth cycle.

From these values we can infer that after the tax cut was implemented the amount of cigarettes increased between the first and second cycle, while continuing to decrease in the third and forth cycles. Also, the rate of decrease between the second and third cycle was lower compared to the rate of decrease between the third and forth cycle.

Comparing the age and sex of those surveyed

In addition to the total number of cigarettes smoked each cycles it is also important to understand the demographics of the survey.

The following table (labeled Table_6) displays numerical summaries of the sex, age and income of those who participated in the survey.

Table 6: The trimmed mean was trimmed by 10 percent

Category	q1	Med	q3	IQR	Mean	trimMean	var	sd
sex	1	2	2	1	1.525	1.531	0.25	0.50
age	2	3	3	1	2.723	2.779	0.75	0.87
income	1	2	3	2	1.957	1.848	0.91	0.95

As a reminder of the values and how they coincide with the data:

- SEX - The sex of the participant:
  - 1: Male
  - 2: Female
- AGEGP4 - The age interval of the participant:
  - 1: 15-19 years old
  - 2: 20-24 years old
  - 3: 25-64 years old
  - 4: 65+ years old
- C1INCAD - The income interval of the participant:
  - 1: Lower class
  - 2: Lower-middle class
  - 3: Upper-middle class
  - 4: Upper class

From (Table_6) we can compare the sex, age and income of those who participated in the survey; however, it is important to note that this data only includes entries of those who gave valid answers to all the survey questions used in this analysis. Those who refused or chose to skip the survey questions were not included in (Table_6). From the table we see that while relatively equal the survey has a slightly higher number of females compared to men with an average trimmed mean being 1.531 and the mean being 1.525. From the quartiles we can see that the median value is 2 also suggesting that the the data contains more females than males and a variance of .25.

When comparing the age of the participants in the data we see that the mean is 2.723 meaning that the most of the people are within the ages of 25-64, which falls in line with current day records where the average age of Canadian smokers is also between 25-64 (see citation (2)). Due to the fact that the value of q1 is 2 we can infer that there a small amount of people in the data who are aged between 15-19, which also makes sense as by Canadian law it is sell cigarettes to those under 19. Lastly, since q3 and median are both 3, coupled with the mean value we can surmise that the majority of people in the data are between the ages 25-64 years old.

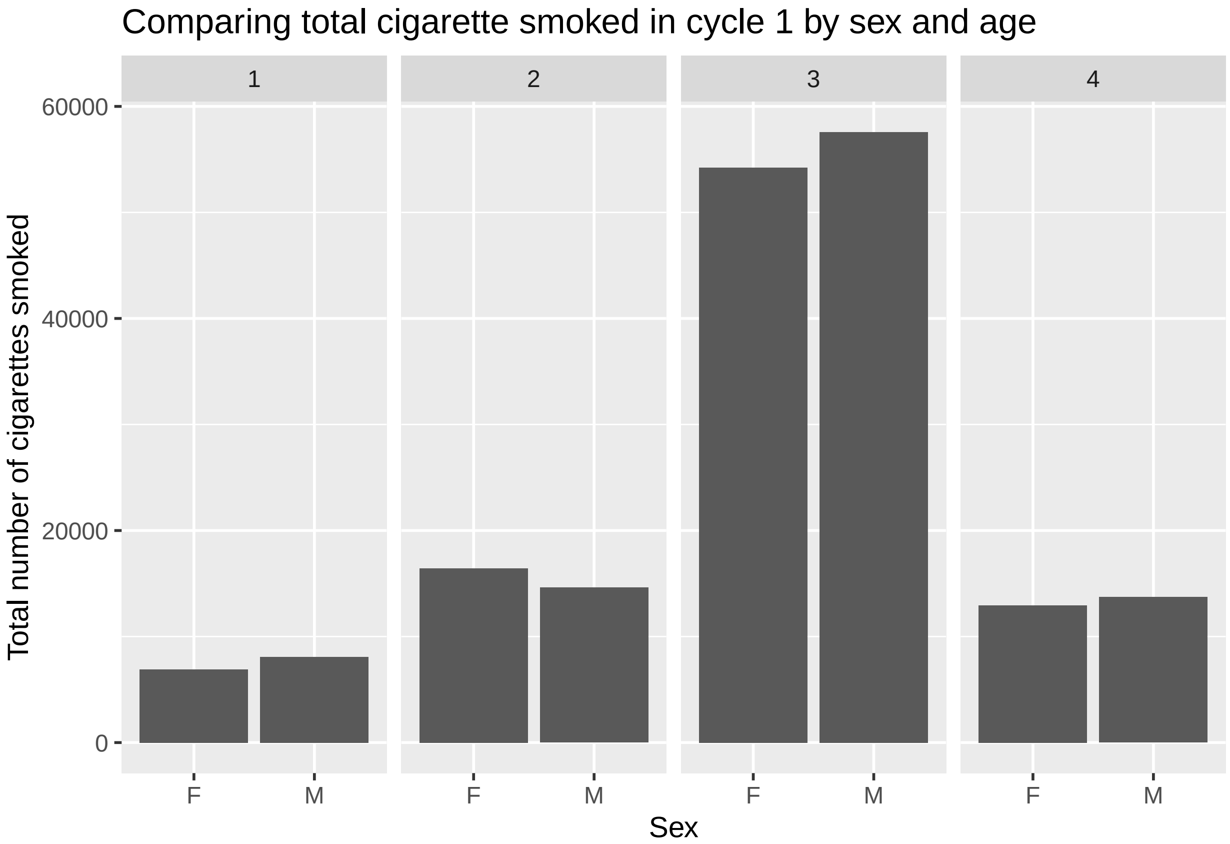
In addition, when comparing the income of those contained in that data we see that from the mean and the trimmed mean that the average income of those who participated in the survey is lower middle class. Furthermore, from the fact the fact that the value of q1 is 1, the value of the median is 2 and the value of q3 is 3, we know that the data contains a large spread of people with the incomes from lower class to upper-middle class with the variance of the income being 0.91. The least represented income in the data is the upper class.

Graphical summaries

In addition to numerical summaries it is also helpful to use graphical summaries to be able to see visual trends in our dataset. To identify the effect that the decrease in tax had on the smoking behaviors of Canadians we can compare the graphs of the total amount of cigarettes smoked by the survey participants in cycles one and four separated by their sex and age to identify any changes to their smoking habits.

Comparing sex and age to the total number of cigarettes smoked in cycle 1

(Graph_1)

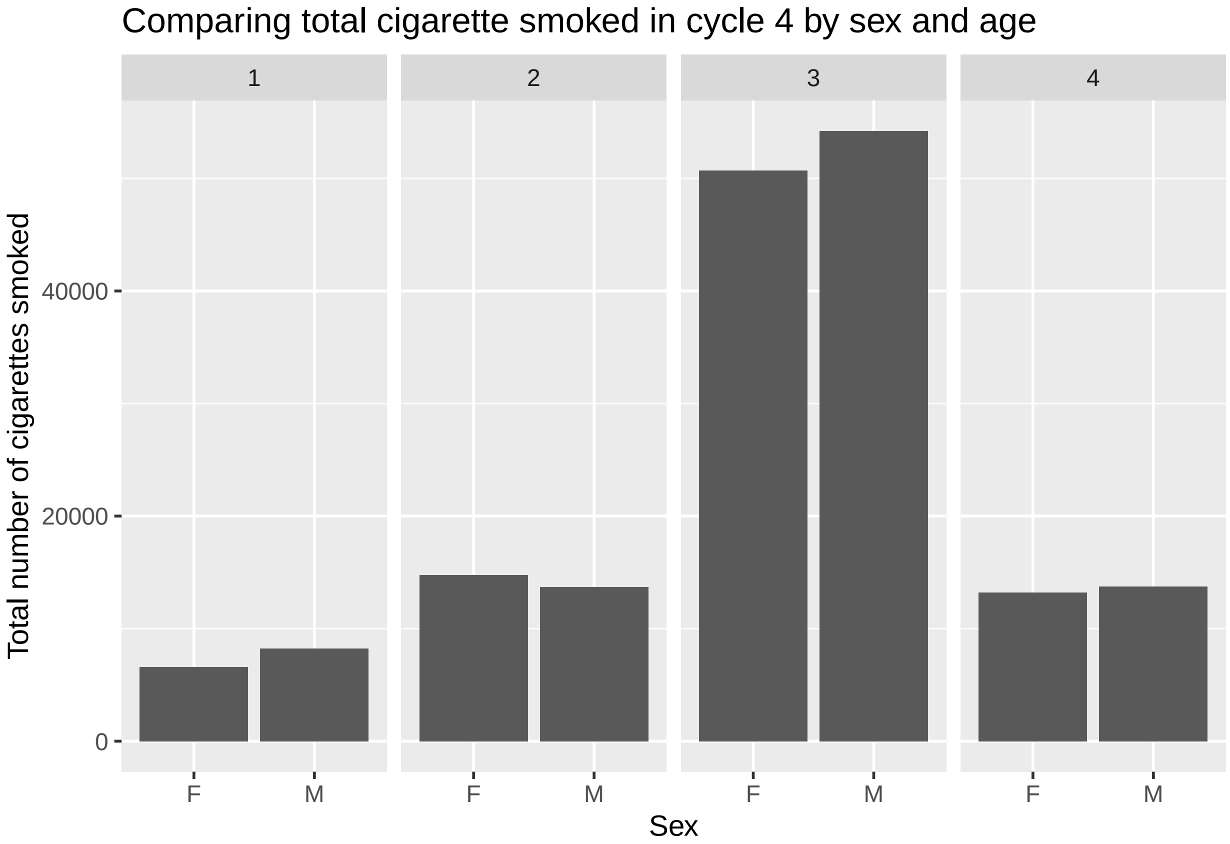


From (Graph_1) we are able see the total amount of cigarettes smoked by the participants in the survey separated by their sex and age where: 1 indicates 15-19 years old, 2 indicates 20-24 years old, 3 indicates 25-64 years old and 4 indicates 65+ years old in the first cycle of questions.

From the graph we see that the 20-24 age interval is the only age interval where the women in the survey smoked more cigarettes compared to the men. In every other age interval we see that men smoked a higher amount of cigarettes compared to the women. In addition, we see that those in the 25-64 years old age interval have a significantly higher total number of cigarettes smoked for both sexes compared to all other age intervals, while those in-between the ages 15 and 19 have the smallest amount of total cigarettes smoked. This makes intuitive sense since the numerical summaries section showed that most individuals in the survey are between the ages 25 and 64. It is worth noting that this age range has the largest range of ages compared to the other three age intervals. Furthermore, it makes sense that those between the ages 15 to 19 have the smallest amount of total cigarettes smoked as they have the lowest population in the survey. Lastly, it is important to remember that any survey participant that chose to skip or refuse to answer a question was not included in the data.

Comparing sex and age to the total number of cigarettes smoked in cycle 4

(Graph_2)



From (Graph_2) we are able the total amount of cigarettes smoked by the participants in the survey separated by their sex and age where: 1 indicates 15-19 years old, 2 indicates 20-24 years old, 3 indicates 25-64 years old and 4 indicates 65+ years old in the forth cycle of questions.

We can see that Graph_2 mirrors many of the trends that were present in Graph_1. In cycle 4 similar to cycle 1 the interval of 20-24 years old is the only age interval where women in the survey smoked more cigarettes in total compared to the men. In addition, also similarly to cycle 1 those between the ages 25 to 64 smoked significantly more cigarettes compared to all other age intervals. However, when comparing the graphs from the first and forth cycles we see that for all age intervals and for both sexes the total number of cigarettes smoked in cycle 4 is lower than the total number of cigarettes smoked in cycle 1. From this we know that the data shows that from February 1994 to February 1995 the total amount of smoking of the participants of the survey decreased.

Resources

This entire document and all its contents were made by using the programming language R and Rstudio. From pulling the data, to cleaning the data, manipulating the data and creating tables and graphs, was all done using R and Rstudio.

Methods

With our data cleaned and analyzed through numerical and graphical summaries we can move onto using the data in models to answer the underlying question of this report on how the effect that decreasing taxes on cigarettes has on the behavior and habits of those who participated in survey. To do this will be using Linear regression modeling, Frequentest Modeling and Bayesian modeling.

Linear regression: Comparing the total amount of cigarettes smoked in each cycle

Using linear regression we will investigate how the total number of cigarette smoked in each cycle changed throughout the year.

We are able to use linear regression modeling to identify trends and fluctuations in the total number of cigarettes smoked in all 4 cycles. To do this we need to graph the number of cigarettes smoked in each cycle and use linear regression to estimate a line of best fit The value of the slope of the line of best fit will help us understand the trends of the behaviors of the participants of the survey in all 4 cycles. In order to do this we will be using the model:

The Model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2)$$

Within this model Y_i is the total number of cigarettes smoked in the ith cycle and x_i is the ith cycle number. In addition the value of β_1 is what we are estimating in the model and it represents the line of best fit which will indicate the trends of the total number of cigarettes smoked in all 4 cycles. $\epsilon_i \sim N(0, \sigma^2)$ represents the component of randomness of the linear relationship between x_i and Y_i. Finally β_0, represents the y intercept of the model in the case that x_i is equal to 0.

Methodologies of the model

Form the data we have cleaned we are able to find the total number of cigarettes smoked in each cycle. Then with the linear regression model we are able to estimate β_1 which can be interpreted as the line of best fit. This estimation will help us understand the if the total number of cigarettes smoked in all the cycles increased or decreased and to what degree.

Assumptions

With the use of linear regression we need to assume the following about our data:

- Independence: Normality: Y_i is distributed normally for any fixed value of x_i,
- Linearity: The relationship the mean of Y_i and between x_i is linear.
- Independence: All observations are independent.
- Homoscedasticity: The variance of residual is the same for any value of x_i.

Data Aggregation

Table 7: Total number of cigarettes smoked in each cycle

Cycle_Number	Total_number_of_cigarettes_consumed
1	184528
2	185416
3	181223
4	175149

From (Table_7) we are able to the total number of cigarettes smoked in each cycle reported in the 1994 survey. With this data we are able to find the line of best fit to better understand how the behaviors and habits of the survey participants were influenced by the decrease in tax of cigarettes. From the data we can see that compared to cycle 1, the total number of cigarettes smoked in cycle 2 showed an increase in number. However, from what we can see in cycle 3 the total number of cigarettes smoked is lower compared to the first 2 cycles and continuing this trend the number of cigarettes smoked in the 4th cycle is the lowest of all cycles.

Frequentist model: Comparing rate of the number of cigarettes smoked in a day for each cycle

Using frequentist modeling we will examine how the average amount of cigarettes smoked in one day has changed throughout the year.

By using frequentist modeling we are able to estimate the average number of cigarettes smoked by each participant in the survey in a day for each cycle. In order to find this value we need to use the following model.

$$X_{u,v} \sim Pois(\lambda_v)$$

For this frequentist model $X_{u,v}$ represents a random variable denoting the number of cigarettes smoked in the cycle u. Furthermore, $Pois(\lambda_v)$ represents an estimate of the average number of cigarettes smoked in a day by a participant in the survey in the cycle u.

Methodologies

From the survey that was conducted by Stats Canada, the number of cigarettes smoked in each day for a week was recorded ($X_{u,v}$) for every cycle. We are able to use that data and find the value of the maximum likelihood estimator of the model. Due to the fact that we are using the frequentist model the value of our parameter of interest ($\hat{\lambda_v}$), is an unknown numerical constant. By finding the parameter of interest's value in each cycle we are able to compare the average number of cigarettes smoked per day between each cycle to better understand the effects that the lowered tax rate had on the smoking habits of the survey participants.

In addition, for the model it was appropriate the use the Poisson distribution due to the fact that we are measuring a number of events (number of cigarettes smoked in a day) in a time interval (in this case the rate is one day) and for the reason that each event is discrete.

Assumptions

As with any models its important to account for the assumptions that have to be made in the model, in this case the source of the assumption lies with the fact that we are using the Poisson distribution family. The assumptions for the model are the following:

- Two or more events can not occur together/simultaneously.
- Events that occur in different intervals of measure are independent from one another
- Expected number of events in each time interval is constant

Parameter of interest

With this model the parameter of interest, $\hat{\lambda}_v$ is the maximum likelihood estimate of our model. By using the data with the maximum likelihood estimate function we are able to find the average number of cigarettes smoked in a day for each cycle u.

Aggregation of data

Table 8: Table contains only first 6 of 10962 rows

daily_c1	daily_c2	daily_c3	daily_c4
1	0	5	3
3	6	4	4
1	6	5	4
8	10	7	8
1	2	2	2
0	0	2	0

From (Table_8) we can see the number of cigarettes smoked by the participants of the survey in one day for each cycle. As the survey accounts for data for all seven days of week and for many different different participants there were 10962 rows of data total. To keep the data table from being too lengthy only the first 6 rows were displayed. In the calculations all 10962 rows will be used. With this data, we are able to find the value of $\hat{\lambda}_v$ for each cycle u which will indicate the average number of cigarettes smoked by the participants of the survey in each cycle.

Hypothesis Test - 2 sided test

With the results of our data we are able to estimate the average number of cigarettes smoked in a day for each cycle. Using this we are able to use hypothesis testing to see if there is any significant changes between the average number of cigarettes smoked in a day between the first cycle compared to the forth cycle. Using this information we will be able to understand how the average number of cigarettes smoked in a day changed throughout the year for those who participated in the 1994 survey.

In the hypothesis test our null hypothesis and our alternative hypothesis are the following.

Null hypothesis (H_0): $\mu_x = \mu_y$ where μ_x is the true population average of the number of cigarettes smoked in a day in the first cycle and μ_y is the true population average number of cigarettes smoked in a day in the forth cycle. This hypothesis states that there was no change in the true population average number of cigarettes smoked in a day between the cycles.

Alternative hypothesis (H_A): $\mu_x \neq \mu_y$ where μ_x is the true population average number of cigarettes smoked in a day in the first cycle and μ_x is the true population average of the the number of cigarettes smoked in a day in the forth cycle. This hypothesis states that there was a significant change in the true population average number of cigarettes smoked in a day between the cycles.

Bayesian model: Comparing rate of the number of cigarettes smoked in a week for each cycle

Using Bayesian modeling we will examine how the average amount of cigarettes smoked in a week day changes from the first to the fourth cycle.

By using Bayesian modeling we are able to estimate the average number of cigarettes smoked by each participant in the survey in a week for each cycle. In order to find this value we need to use the following model.

$$likelihood : X_{1,u}, ..., X_{n,u} \sim Pois(\lambda_u)$$

In order to find the average number of cigarettes smoked in a week for cycle u in the Bayesian model, the Poisson distribution is used for the likelihood of the model. It is appropriate to use the Poisson distribution due to the reason that we are trying to find the average of a discrete number of events that occur within a certain time interval, in this case we are trying to find the average number of cigarettes smoked (discrete) in one week. So for our model, for the data in cycle u, the average number of cigarettes smoked (discrete) in one week will follow $Pois(\lambda_u)$.

$$prior : \lambda_u \sim gamma(\alpha = 5, \beta = 25)$$

Within the Bayesian model, the prior utilizes the gamma distribution family. The gamma distribution was chosen as it allows for the posterior distribution to be in an recognizable form and the fact that in this circumstance it would cause λ_u to be defined to be greater or equal to 0. This is appropriate as the number of cigarettes smoked has to be equal or greater than 0. In order to find the values of alpha and beta, a study from Waterloo was referenced (see citation (1)). The Waterloo study contains the average amount of cigarettes smoked in a day from the years 1999 to 2017, through the use of finding a line of best fit and extrapolating the data it was found that the average number of cigarettes smoked in 1994 according to the Waterloo study data is 18.4. From there by multiplying by 7, it was further extrapolated that the average number of cigarettes smoked in a week is 128.8. Using 128.8 as our mean value it can be inferred that the range of possible values of the average number of cigarettes smoked in a week is between in interval of 0 to 257.6, with 128.8 being the mean value. Lastly, it makes intuitive sense that the gamma distribution PDF should be right skewed to account for heavy smokers who smoke significantly more than the average amount of cigarettes. From this information I chose alpha to be 5 and beta to be 25 as the PDF of the beta distribution with those parameters contains the aforementioned qualities with a similar mean, range of values and is right skewed.

$$posterior : gamma(\sum x_{i,u} + \alpha, n_u + \beta)$$

With the prior and likelihood it was found that the post posterior is $gamma(\sum x_{i,u} + \alpha, n_u + \beta)$ for the cycle u.

Methodologies

From the survey that was conducted by Stats Canada, the number of cigarettes smoked in each day in a week was recorded for all cycles and from that data we are able to find the number of cigarettes smoked in a week. From this data, we are able to find the value of alpha and beta in our gamma posterior distribution and the mean the gamma posterior represents the value of the average amount of cigarettes smoked in a week.

Assumptions

As with any model it's important to account for the assumptions that have to be made in the model, in this case the source of the assumption lies with the fact that under a Bayesian model we have a Poisson likelihood and a gamma prior.

- Two or more events can not occur together/simultaneously.
- Events that occur in different intervals of measure are independent from one another
- Expected number of events in each time interval is constant
- For the gamma distribution alpha and beta must be greater than zero.
- Since $gamma(\alpha, \beta)$ can only result in a values of zero or greater we assume that λ_u is also always greater or equal to zero.

Parameter of interest

Due to the fact that we are using a Bayesian model through the use of the posterior and data we are able to find the mean of the posterior gamma distribution for all cycles. The mean of the gamma posterior gives us the value of the parameter of interest which is $\hat{\lambda}_{u,bayes}$. The value of $\hat{\lambda}_{u,bayes}$ represents the average number of cigarettes smoked in a week for cycle u.

Aggregation of data

Table 9: Total number of cigarettes smoked in a week for each cycle

Cycle_1_total	Cycle_2_total	Cycle_3_total	Cycle_4_total
5	17	39	17
24	34	40	36
12	41	33	28
56	70	49	56
12	16	20	14
31	35	12	47

From (Table_9) we can see the number of cigarettes smoked by the participants of the survey in a week for all four cycles. As the survey accounts for data for many different participants there were 1566 rows of data total so to keep the data table from being too lengthy only the first 6 rows were shown. In the calculations all 1566 rows will be used. With this data, we are able to find the value of the of the average number of cigarettes smoked in a week by the participants of the survey in each cycle.

Hypothesis Test - 2 sided test

With the results of our data we are able to estimate the average number of cigarettes smoked in a week for each cycle. Using this we are able to use hypothesis testing to see if there is any significant changes between the average number of cigarettes smoked in a week between the first cycle compared to the forth cycle. Using this information we will be able to understand how the average number of cigarettes smoked in a week changed throughout the year for those who participated in the 1994 survey.

In the hypothesis test our null hypothesis and our alternative hypothesis are the following.

Null hypothesis (H_0): $\mu_a = \mu_b$ where μ_a is the true population average number of cigarettes smoked in a week in the first cycle and μ_b is the true population average number of cigarettes smoked in a week in the forth cycle. This hypothesis states that there was no change in the true population average number of cigarettes smoked in a week between the cycles.

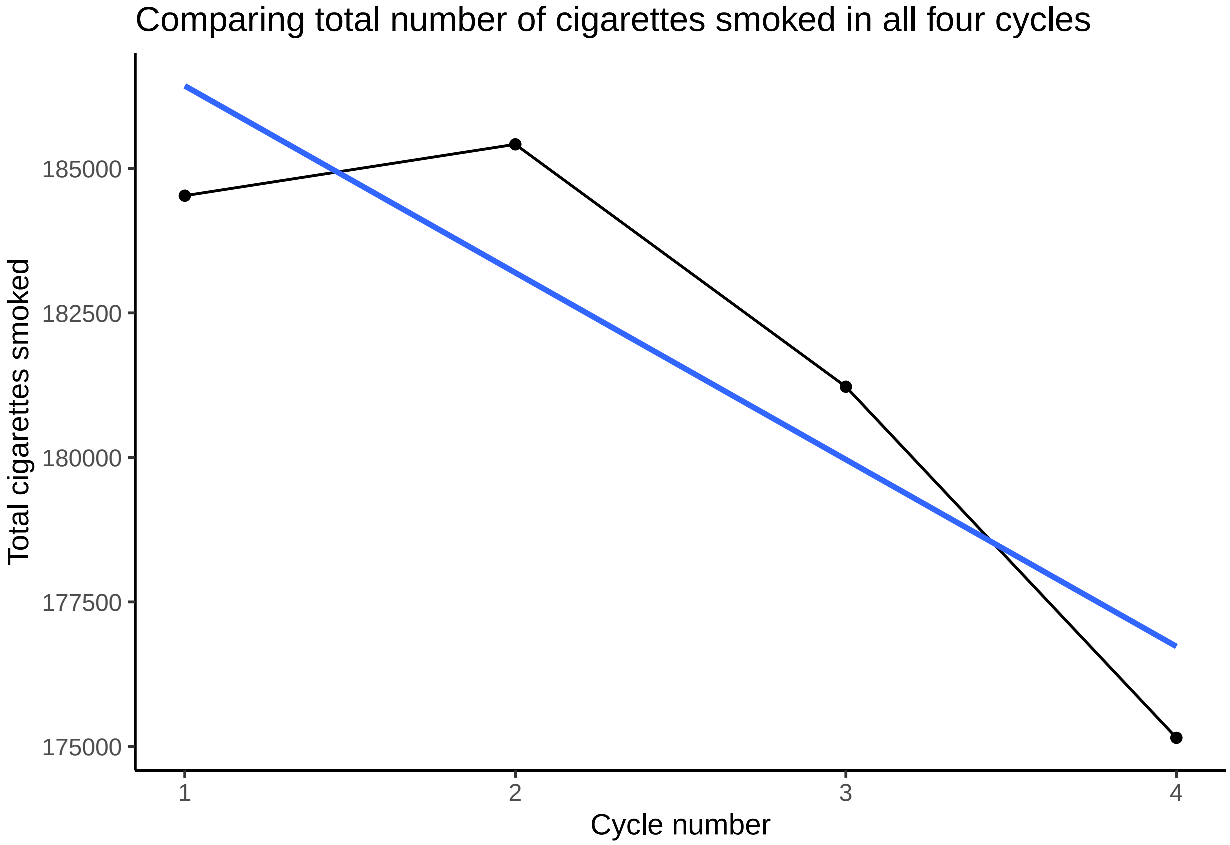
Alternative hypothesis (H_A): $\mu_a \neq \mu_b$ where μ_a is the true population average number of cigarettes smoked in a week in the first cycle and μ_b is the true population average number of cigarettes smoked in a week in the forth cycle. This hypothesis states that there was a significant change in the true population average number of cigarettes smoked in a week between the cycles.

Results

Results: Comparing the total amount of cigarettes smoked in each cycle

Using the data from (Table_7) we are able to use linear regression to estimate the value of β_1 which represents the line of best fit which will indicate to us what the trends are of the total number of cigarettes smoked in all four cycles. Using our data the result of β_1 is -3233 which means that on average between all four cycles the total number of cigarettes smoked per cycle decreased an average of 3233 cigarettes. To gain a better understanding we can graph the data.

(Graph_3)



In (Graph_3) we are able to see the total number of of cigarettes smoked in all four cycles with the black line. The blue line represents the value of β_1 which represents that line of best fit which resulted in being -3233. From this we are can surmise that on average per cycle the total number of cigarettes smoked decreased by 3233 cigarettes.

Interpreting of the results

From the results (graph_3) which contains the linear regression results we are able to investigate how the total number of cigarette smoked in each cycle has changed throughout the year. This information will help us better understand how the the decrease in tax of cigarettes sales influenced the habits of the participants in the 1994 survey.

From the graph we see that compared to the first cycle, the total number of cigarettes smoked in the second cycle was higher than the total amount of cigarettes smoked in the first cycle. However, the total amount of cigarettes smoked in the third cycle was lower than both of the first two cycles and the total amount cigarettes smoked in the forth cycle resulted in the lowest amount of cigarettes smoked from all four cycles. Furthermore, it is important to note that the steepest slope is the decreasing slope from cycle three to four, the second steepest slope is the decreasing slope from cycle two to three and the least steep slope is the increasing slope from cycle one to two. Due to this, the value of β_1 resulted in a negative slope which represents the trend that for every cycle the total number of cigarettes smoked decreased on average of 3233 cigarettes per cycle. From this we can infer that when the when the cigarette tax was first introduced, it caused an increase on the total amount of cigarettes purchased and smoked, but by cycle three the total amount of cigarettes purchased and smoked decreased compared to the first two cycles and finally in the forth cycle the number of cigarettes purchased and smoked decreased by an even greater amount compared to the amount reported in cycle three. To understand why the total amount of cigarettes smoked decreased despite the decrease in tax, you need to compare the total amount of cigarettes smoked from from 1965 to 1994, from a study done by the University of Waterloo (see citation (1)) we see that that from 1965 to 1994, the smoking prevalence in Canada has been on a historical downward trend going from 50 percent of adult Canadians smoking to 35%. As our knowledge on the dangers of smoking increasing there has been an increase in public awareness campaigns designed to inform the public about the cancer risks and how to quit smoking. As the number of cigarettes smoked was on a downward trend we that from the data that the decrease on taxes slowed this historic drop in smoking with as we see an increase in the total number of cigarettes smoked compared to cycle 1 and less steep slope of the total amount of cigarettes smoked from cycle three to two compared to cycles four to three. In addtion the Canadian provinces that did not decrease the tax on cigarettes had a larger decrease in the prevelance of smoking compared to all instances of this survey (see citation 5). Moreover, we from what we can see the results do seem reasonable. As we learn more about smoking and the health risk associated with smoking there has been many public awareness campaigns and resources to help people quick smoking. We see from the Waterloo smoking study (see citation (1)) that from 1965 to 2017 the smoking prevalence in adults has decreased from 50 to 15 percent. In 1994 with the decrease in the cigarette tax, as cigarettes became cheaper and more accessible, it makes sense that the drop in smoking rates saw a decrease.

Results: Comparing rate of the number of cigarettes smoked in a day for each cycle

From the data observed in (Table_9) we are able to find the value of the parameter of interest ($\hat{\lambda}_v$) for each cycle by using our data in conjunction with the maximum likelihood estimator of our model. To see the full derivation of the maximum likelihood estimator please see the appendix. From this value we can determine the average number of cigarettes smoked in one day by the participants of the survey for all four cycles. In addition, we can also find the 95 percent confidence interval of value of ($\hat{\lambda}_v$) in each cycle by using the equation $\hat{\lambda}_v \pm 1.96 * \sqrt{\hat{\lambda}_v/n}$ (see citation (3)) where n is the total number of observations and $\hat{\lambda}_v$ is the maximum likelihood estimator value of the average amount of cigarettes smoked in one day.

Table 10: Average number cigarettes smoked in one day for all cycles

cycle	daily_avg	Confidence_interval_95_percent
Cycle_1	16.83342	(16.757,16.910)
Cycle_2	16.91443	(16.837,16.991)
Cycle_3	16.53193	(16.456,16.608)
Cycle_4	15.97783	(15.903, 16.053)

From (Table_10) we are able to see the average number of cigarettes smoked in one day for all four cycles. In addition we also see the confidence intervals for the daily average number of cigarettes smoked in one day for all four cycles. With the confidence intervals we can say that we are 95 percent confident that for each cycle the interval given will include true population paramter λ_v.

Interpreting of the results

From the results (Table_10) which contains the frequentist model results we are able analyze how the average amount of cigarettes smoked in one day has changed between the first and forth cycle. This information will help us better understand how the the decrease in tax of cigarettes sales influenced the habits of the participants of the 1994 survey.

From the (Table_10) we see that the results for the frequentist model mirror the results from the linear regression model. The average of the number of cigarettes smoked in one day in cycle two increase compared to cycle one, while the average number of cigarettes smoked in day in cycle three are lower than bother cycle one and two and the daily average number of cigarettes smoked in cycle four are lower than all other cycles. Notably the magnitude of decrease from cycle three to four is higher than the magnitude of decrease from cycle two to three and the magnitude of increase from cycle one to two. As we have seen from the University of Waterloo study, (see citation 1) the smoking prevalence of Canadians has greatly decreased from 1965 to 1994, decreasing from 50 percent of Canadians adults smoking to 35 percent. The data seen in (Table_10) reinforces the idea that as cigarettes tax decrease, it slows down the fall in the smoking prevalence of Canadians by making cigarettes more accessibly cheaper to purchase. We know that these results are intuitively reasonable, for similar reasons we knew that the linear regression results were reasonable. Due to the dangers that cigarettes pose to our health there have been many public campaigns and avenues to allow for smokers to quit smoking to be constructed. This has caused a continual decrease in the prevalence of smokers in Canada; however, in 1994 as cigarettes became cheaper to buy this incentivised smokers to increase their daily average number of cigarettes as cigarettes became more affordable. This caused the trend of decreasing prevalence of smokers in Canada to slow down. It is important to note that provinces that did not decrease cigarette tax rates had an even larger decrease in the prevalence of smokers than what we have seen from cycle three to cycle four in this survey (see citation 5).

Hypothesis Test Results

Using the t.test function in R it was found that the p value for the significance in the fluctuation between the average number of cigarettes smoked in cycle 1 and cycle 4 was 3.976e-11. From this value we can infer that since the p value is small, that there was a significant change in the average number of cigarettes smoked from cycle 1 to 4. From this we can accept out alternative hypothesis. This makes intuitive sense due to the reason that as we can see from the university of Waterloo study (see citation 1) that due to public awareness about the dangers of smoking the prevalence in smoking has been dropping in Canadian adults. Due to the effects of decreasing the tax on cigarettes the rate of decrease in the prevalence of smoking has slowed down, but from cycle one through cycle four enough people decreased their smoking habits for our alternative hypothesis to be accepted. We can infer intuitively that without the decrease on cigarettes tax the decrease of the average number of cigarettes smoked in one day would be even lower between cycles one and four causing an even smaller p value.

Results: Comparing rate of the number of cigarettes smoked in a week for each cycle

Using the data from (Table_9) we are able to use our data to find the value of our posterior for all cycles and then through the mean of the posterior we will be able to find the average amount of cigarettes smoked within a week for all cycles. We can use this data to better understand how the decrease in the sales tax of cigarettes effected the behavior of those who participated in the survey.

Table 11: Average number cigarettes smoked in a week for all cycles

cycle_number	daily_avg2	Credible_interval_95_percent
Cycle_1	115.986	(115.457,116.515)
Cycle_2	116.544	(116.014,117.075)
Cycle_3	113.908	(113.384,114.433)
Cycle_4	110.091	(109.576, 110.607)

In (Table_11) for all four cycles by finding the mean of the posterior we have found the estimated average of the number of cigarettes smoked in a week. In addition for the estimates there is also included a 95 percent credible interval. For the credit interval there is a 95 percent probability that the true mean value of the average amount of cigarettes smoked in a week is in the interval in (Table_11).

Interpreting of the results

From the results (Table_11) which contains the Bayesian model results we are able analyze how the average amount of cigarettes smoked in a week has changed between the first and forth cycle. This information will help us better understand how the the decrease in tax of cigarettes sales influenced the habits of the participants of the 1994 survey.

The values in (Table_11) reinforce what we have found within the finding in our linear regression model and frequentist model. Similar to the other two models, we an increase in the average number of cigarettes smoked from cycle one to cycle two, but a decrease in the the average number of cigarettes smoked in a week from cycle two to cycle three and from cycle three to cycle four. Also mirroring the other two models we see that the decrease in the average number of cigarettes smoked in a week from cycle three to four is steeper than the decrease from cycle two to three and the increase from cycle one to two. From this information we can infer that when the tax on cigarettes decreased it caused the participants of the survey to increase their smoking rates, cause an uptick in the average number of cigarettes smoked in a week. However, after from the second cycle onward the participants on average decreased their smoking habits on average. This data is reasonable as it from the study from the university of Waterloo (see citation 1) on smoking rates in Canada, from 1965 to 1994 there has been continuous decline in the prevalence in smoking for Canadians. This is mainly due to our increasing understand of the dangers of smoking and the awareness campaigns to decrease smoking rates. As we have seen from the data, when the tax decreased, it caused an increase in the smoking habits of the survey participants. When cigarettes become cheaper they become more accessible causing a slowing down of the decreasing rates smoking prevalence in Canada.

Hypothesis Test Results

From our hypothesis test, using the t.test function in R we have found the p value to be 0.009765. Since the P value is so small we reject the null hypothesis and accept the alternative hypothesis. This means that there was a significant change in the the average number of cigarettes smoked in a week when comparing cycles one and four. The significant change was a decrease in the average number of cigarettes smoked from cycle one to cycle four. The decrease is due to people slowing down their smoking habits, due to the cancer risks. The decrease in the cigarettes sales tax caused a slow in the in the decrease in the rates of smoking, and we can infer that without the drop in tax the decrease in the average number of cigarettes smoked in a week when comparing cycles one and four would be greater.

Conclusion

In the year 1994 the government decreased the tax rate of cigarettes in some provinces to curb the selling of illegal cigarettes. To investigate the impact that a lower tax rate on cigarettes would have on smokers Health Canada asked Stats Canada to conduct a year long survey with participants who were established smokers. The survey began in February 1994 and concluded in February 1995. Within the survey, the same group of participants were asked the same set of questions about their smoking habits every three months. With the data from the survey, the goal of this report is to investigate what effect the decrease on the taxes of cigarettes had on the smoking habits and behaviors of the participants of the survey. In order to figure this out, we used a Linear regression model to analyze how the total number of cigarettes smoked fluctuated between all four cycles. In addition, we used a frequentist model with a Poisson distribution family to investigate how the average number of cigarettes smoked in a day changed throughout the four cycles. Lastly we used a Bayesian model with a Poisson distribution family for the likelihood and Gamma distribution family for the prior in order to analyze how the survey participant's average number of cigarettes smoked in a week changed throughout the four cycles. The results of the three models reinforced each other. In each model it resulted in an increase in the number of cigarettes smoked from cycle one to cycle two, but resulted in an decrease in the number of cigarettes smoked from cycle two to cycle three and from cycle three to cycle four. In all cases the slope of the decrease from cycle three to cycle four was steeper than both the decreasing slope from cycles two to three and the increase from cycles one to two. This caused an overall decrease to the average number of cigarettes smoked in a day, a decrease in the average number of cigarettes smoked in a week and a decrease in the total number of cigarettes smoked by all participants in the survey when comparing the results from cycles one and four. We see in the Linear regression model that the value of β_1 was -3233 meaning that on average in each cycles the total amount of cigarettes smoked by all the participants was decreasing by 3233 cigarettes. Recall the blue line from (graph_3) From the result of our hypothesis test in our frequentist model and Bayesian model, the small p value in both cases also indicate to us that the decrease in the average number of cigarettes smoked in a week and day were both significant.

To explain the decrease in smoking rates, from the study of University of Waterloo (see citation (1)) we see that from 1965 to 1994 the prevalence of smoking adults dropped from 50 percent to 30 percent. As our scientific knowledge on the dangers of smoking cigarettes increased it caused many people to cut down or give up on smoking. There have been many public awareness campaigns in schools and in the media to inform people on the cancer risks and many new avenues have been researched on how to help people quit cigarettes. Due to this we see that more and more people are giving up smoking which caused a decreased in the smoking rates in the data. However, while the prevalence of smoking was decreasing year by year, we see that from cycle one to cycle two the total number of cigarettes and average number of cigarettes smoked in a day and week increased. In addition, we see that the decreasing slope from cycle three to four is steeper compared to cycles two and three. From this we know that there was an initial uptick in the smoking habits of the survey participants that then started to decrease from cycle two onwards. When cigarettes become cheaper they become more affordable and it gives people an incentive to smoke as it is no longer as expensive to buy cigarettes. So many people who refrained from smoking due to high cost then started smoking. From the data and our inferences we can conclude that while on a whole Canadian smoking prevalence is decreasing, decreasing the taxes on cigarettes causes the rate of decreasing smoking prevalence to slow. The Canadian provinces that did not enact the lowering of cigarettes taxes observed a greater decrease in smoking prevalence compared to those who participated in this survey (see citation (5)).

Within the survey there were numerous drawbacks and limitations. One drawback was that fact many participants had to be filtered out due to them refusing or skipping certain questions in the survey. Furthermore, another drawback is that the survey only included cigarettes and did not consider other tobacco products such as chewing tobacco and electronic cigarettes which is a very popular source of tobacco in 2021. When the price of cigarettes changes it is important to see how other forms of tobacco intake change for tobacco users.

As for next steps, I would recommend that that the survey be re done in as the world today is quite distinct compared to 1994. There are many more avenues for one to use tabacoo. In addition to cigarettes the survey should also include electronic cigarette as a significant number of smokers now use electronic cigarettes instead of traditional cigarettes. Electronic cigarette are especially popular for people in their teen years and is vital to understand the behaviors of current day smokers. Lastly, it is important to provide survey questions and choose survey participants to get the highest amount of valid answers as possible. In that case it would limit the amount of responses that would be filtered out.

Bibliography

(1) Reid JL, Hammond D, Tariq U, Burkhalter R, Rynard VL, Douglas O. “Tobacco Use in Canada: Patterns and Trend’s”, 2019 Edition. Waterloo, ON: Propel Centre for Population Health Impact, University of Waterloo. 2019. Accessed 23 August 2021. https://uwaterloo.ca/tobacco-use-canada/

(2) Gilmore, J. “Report on Smoking Prevalence in Canada”, 1985 to 1999 (Statistics Canada, Catalogue 82F0077XIE). January 2000. Accessed 23 August 2021. https://www150.statcan.gc.ca/n1/pub/82f0077x/4193722-eng.pdf

(3) Dr Deng. “Computing confidence interval for poisson mean. Computing Confidence Interval for Poisson Mean”. March 20 2014. Accessed 23 August 2021. http://onbiostatistics.blogspot.com/2014/03/computing-confidence-interval-for.html.

(4) ODESI2. Scholars Portal. http://odesi2.scholarsportal.info/webview/. Accessed 23 August 2021.

(5) Vivian H Hamilton, Carey Levinton, Yvan St-Pierre, Franque Grimard. “The effect of tobacco tax cuts on cigarette smoking in Canada”. Canadian Medical Association. 15 January 1997. Accessed 23 August 2021. https://untobaccocontrol.org/taxation/e-library/wp-content/uploads/2019/07/Hamilton1997.pdf

Appendix

Calculating Maximum liklihood estimator for Poission distribution.

n = number of observations

$P(Y_i = y_i|\lambda) = \frac{e^{-\lambda} * \lambda^{y_i}}{y_i!}$

$L(\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda} * \lambda^{y_i}}{y_i!}$

$L(\lambda) = \frac{e^{-n\lambda} * \lambda^{\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} y!}$

l(λ) = ln(L(λ)) = -nλ + ($\sum_{i=1}^{n} y_i$)ln(λ) - $\sum_{i=1}^{n} y_i!$

l'(λ) = -n + ($\sum_{i=1}^{n} y_i/\lambda$) = 0

$\hat{\lambda} = \sum_{i=1}^{n} y_i$/n

second derivative test

l"(λ) = -$\sum_{i=1}^{n} y_i/\lambda^2 < 0$

Since λ^2 will always be positive due to the square and the value of the numerator is always negative since all y_i are greater or equal to 0, we know that the equation will always be less than 0. Thus,

the value of MLE of Poisson distribution is $\hat{\lambda} = \sum_{i=1}^{n} y_i$/n