A Statistical Study On Obesity Factors

Eric Liu 1005351717

August 23, 2021

Abstract

This report investigates several possible factors that relate to the prevalence of overweight and obesity in hopes to provide a better understanding on the growing problem of obesity. The data used in this report is taken from the The Canadian Community Health Survey (CCHS) particularly the 2017-2018 annual component, and it encompasses survey information about the health status of Canadians. The data section of this report provides summary statistics on several variables that may be a reasonable factor of obesity and out of these select variables, it was determined that healthy eating and physical activity that had largest difference in their BMI averages. Therefore these variables were tested by a frequentist exponential model and a Bayesian normal model to check the level of statistical significance. The frequentist model covered the physical exercise component and it concluded that on average, individuals with non-overweight BMI experience more physical activity than overweight individuals with approximately 18 more minutes of physical activity. The Bayesian model covered the healthy eating component and it concluded that on average, individuals that consume healthy foods more frequently tend to have have a lower average BMI. In all models, a hypothesis test was done to determine the significance of these factors, and in all cases they were found to be statistically significant though t-testing.

Introduction

Obesity is one of many health issues growing in the world especially in North America and it is a condition where a person accumulates excessive fat that may affect their physical health and limit their performance in everyday life activities. The purpose of this report is to investigate several variables that may have a direct causal relationship with obesity through a variety of statistical models. The data used in this report is the The Canadian Community Health Survey (CCHS) taken from the health section under the Ontario Council of University Libraries (ODESI) and it contains general information about an individual's weight, BMI, and overall health status. The data section will introduce some more important variables to consider and their summary statistics on their mean, median, variance, etc. There are several hypotheses in this report that attribute to the global subject on whether or not obesity is a growing problem and all of these will be discussed into further detail in the models and results section. The first model introduced is a linear regression model which will predict the hypothesis on whether or not the annual body mass index (BMI) has remained constant throughout the years. The next hypothesis involves the topic of physical exercise and if it does in fact have a beneficiary relationship on an individuals BMI. An exponential model will be used to model this case. The final hypothesis will speculate if there is any statistical significance on the relationship between healthy eating and individual BMI, which will be modeled after a Bayesian normal model. Hopefully by the end of this report, the reader is provided with a thorough understanding of the statistical analysis of the data and perhaps gain new insights on the growing problem of obesity.

Data

Data Collection

The Canadian Community Health Survey (CCHS) is created by specialists from Statistics Canada where they survey and collect data on the health of Canadians. Around every one or two years there are approximately 120,000 adult Canadians surveyed. The survey is distributed equally across all regions in Canada relative to their population size to give the best representation all Canadians. The survey that will be used frequently particularly in this report will be the CCHS 2017-2018 annual component. The CCHS administers a variety of questions relating to health status, health care, and health determinants, but for this report the main topic of discussion are the factors attributing to obesity and how much has it changed over the years. Therefore only a subset of the total data is gathered, which include variables such as the weight and BMI of an individual.

Data Cleaning

The data set for this report is a subset of 9 variables chosen from the original CCHS data set. Since the original data set is very large, only a few columns were selected to motivate the topic of discussion and these variables in particular are described below. Furthermore, a quick check on all the individual data values was done which resulted in none of them having any duplicates. All individual data were unique. Another variable created was a Weight (lbs) variable since as the original data was only measured in kilograms.

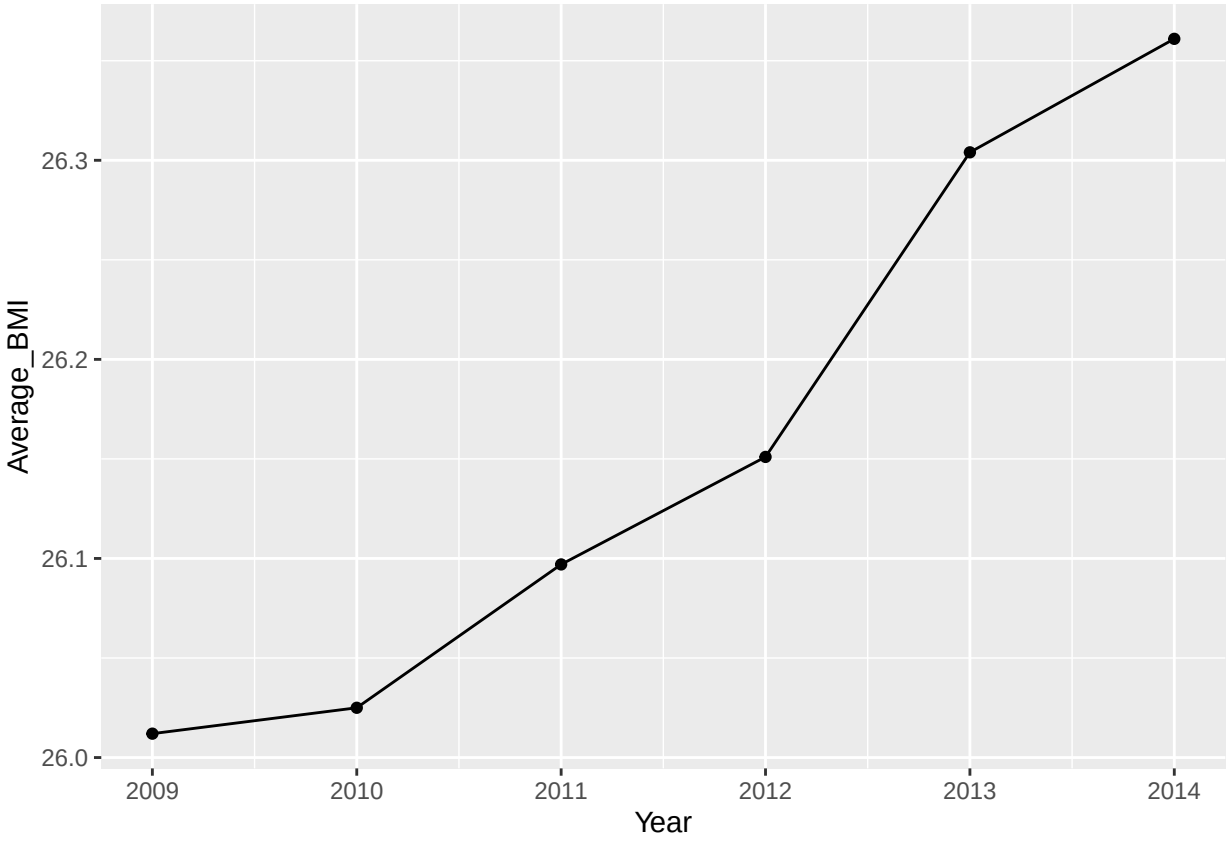
Data Set Variables

- Weight (kg)
- Weight (lbs)
- BMI
  - The Body Mass Index of an individual
    - Underweight is BMI less than 18
    - Normal weight is BMI between 18 to 25
    - Overweight is BMI between 25 to 30
    - Obese is BMI greater than 30
- Age
  - Primarily focused on adults (18+ years)
- Sex
  - Male/Female
- Weekly amount of physical activity
  - Measured by amount of hours of physical activity per week in a sense that an individual has participated in some sort of vigorous exercise or sports activity
- Hours of sleep
- Healthy consumption type
  - How often a person eats fruits and vegetables per day categorized into 3 parts
    - Less than 5 times per day
    - 5 to 10 times per day
    - More than 10 times per day
- Type of drinker
  - How often a person drinks categorized into 3 parts

    * Regular drinker
    * Occasional drinker
    * Non-drinker

- Type of smoker
  - How often a person smokes categorized into 3 parts
    * Daily
    * Occasionally
    * Non-smoker

Graph 1



The points on Graph 1 were taken from ODESI website on the summary statistic section for the survey of that year and they represent the average BMI for each year. The BMI is sometimes not measured every year, so the most consistent group of years chosen were 2009-2014. The graph exhibits an upward trend implying there exists an increase in average BMI per year. Later in this report, a linear regression model will be used estimate this increase and also determine if this change is statistically significant.

(Table 1) Summary statistics for Body Mass Index (BMI)

Sex	Maximum	3rd Quartile	Median	IQR	Mean	TrimmedMean20	SD	Variance
Female	56.72	30.91	26.60	7.68	27.64	26.81	5.92	35.04
Male	57.48	31.08	27.65	6.16	28.30	27.80	5.01	25.07

Age	Maximum	3rd Quartile	Median	IQR	Mean	TrimmedMean20	SD	Variance
Less than 40	57.48	30.91	27.13	6.98	27.87	27.24	5.48	30.06
Older than 40	53.86	31.51	27.53	7.21	28.46	27.74	5.73	32.86

Hours_of_sleep	Maximum	3rd Quartile	Median	IQR	Mean	TrimmedMean20	SD	Variance
Greater than 7	57.48	30.20	26.50	6.68	27.29	26.68	5.21	27.16
Less than 7	56.56	31.08	27.22	7.05	27.94	27.31	5.53	30.58

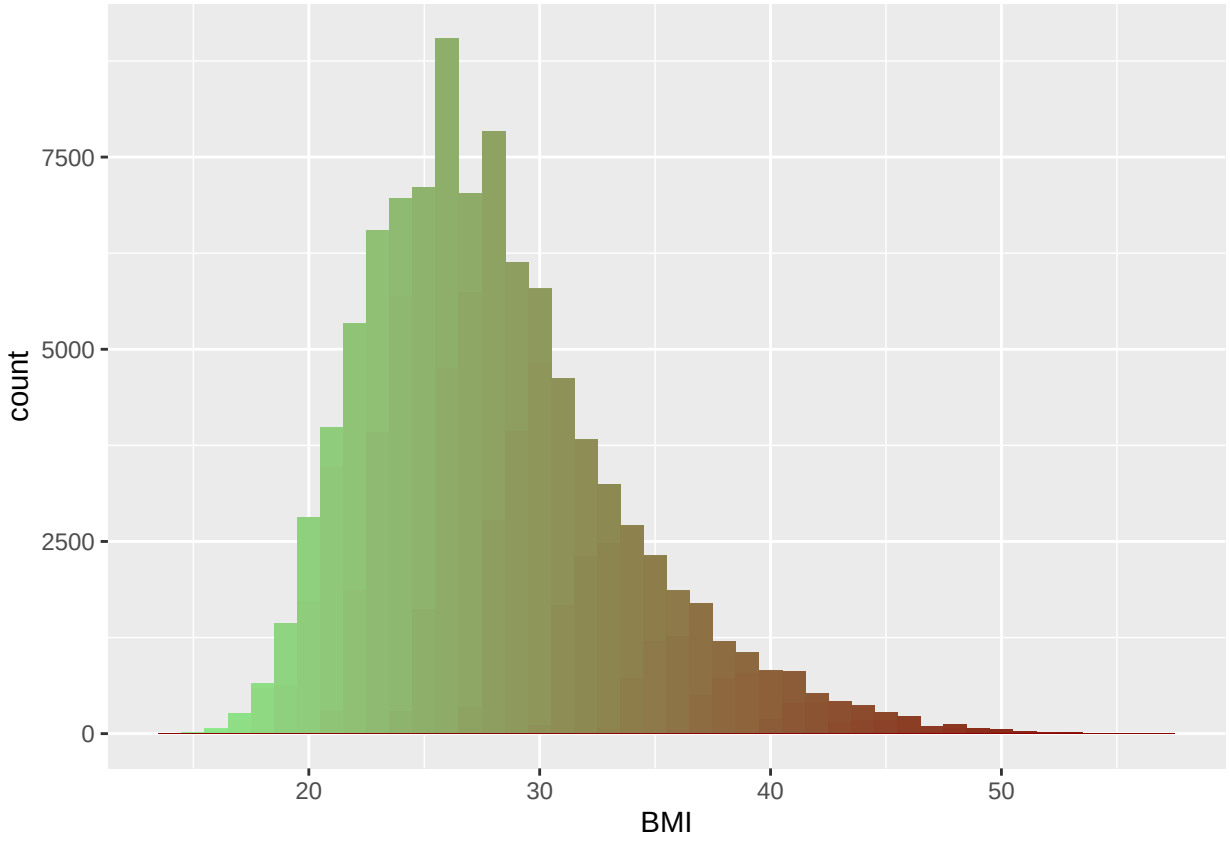
Healthy_consumption_type	Maximum	3rd Quartile	Median	IQR	Mean	TrimmedMean20	SD	Variance
Between 5 to 10 times per day	50.02	31.19	26.77	7.16	28.06	27.17	5.70	32.52
Less than 5 times per day	53.77	32.30	27.64	8.13	28.69	27.96	6.03	36.32
More than than 10 times per day	42.71	30.36	25.75	6.46	27.56	26.52	5.65	31.93

Type_of_smoker	Maximum	3rd Quartile	Median	IQR	Mean	TrimmedMean20	SD	Variance
Daily	53.45	30.58	26.52	7.23	27.48	26.82	5.62	31.63
Not at all	57.48	31.12	27.29	6.95	28.07	27.43	5.51	30.33
Occasionally	56.72	30.20	26.50	6.70	27.30	26.67	5.33	28.45

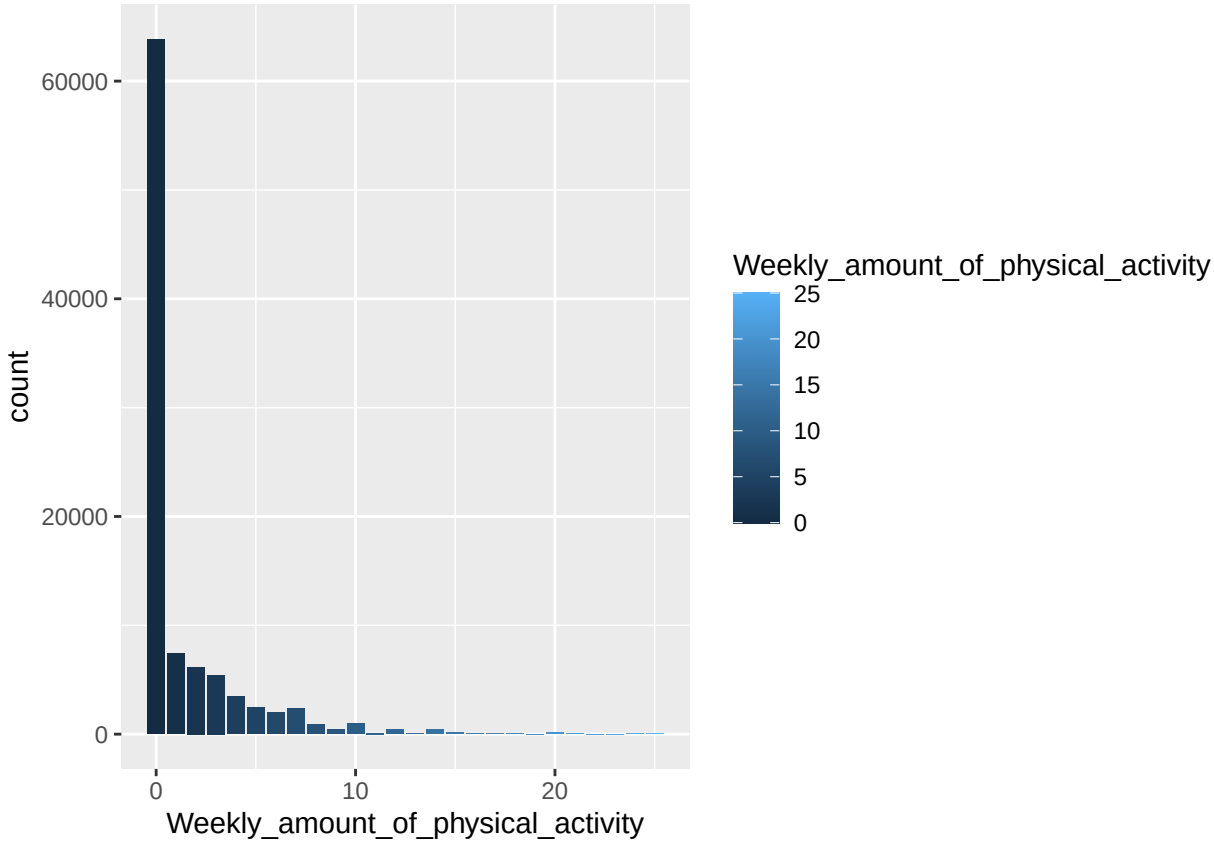
Type_of_drinker	Maximum	3rd Quartile	Median	IQR	Mean	TrimmedMean20	SD	Variance
Daily	56.56	30.58	27.01	6.49	27.73	27.16	5.19	26.98
Not at all	57.48	31.40	27.25	7.61	28.06	27.39	5.90	34.84
Occasionally	56.72	32.21	27.64	8.18	28.63	27.90	6.17	38.04

Table 1 shows the summary statistics for all variables that may relate to or affect an individual's weight. There are several variables that attribute to greater mean BMI such as the decreased amount of sleep, old age, being a non-smoker, but the variable to notice with the most change in the mean and the median BMI is the healthy consumption type. The healthy consumption type refers to the amount of healthy foods (fruits and vegetables particularly) that an individual consumes per day. The difference in BMI differs by nearly 2 in their medians. In all variables the mean is in the realm of 26-28, which is considered overweight (BMI of 25-30) and the 3rd quartile for all variables are above 30, meaning around 25% of all Canadians in this survey are considered obese. The graph below will give a better visual representation of the BMI count proportion.

Graph 2



Graph 2 above shows a histogram that counts all the individual BMIs. The graph closely resembles a normal curve with a possible mean of approximately 27-28. The mean of this graph can also be estimated through modeling under a normal distribution, and then consequently bootstrapping a confidence interval to give a range of values where it is most confident where the mean is.



Graph 3

Graph 3 shows a bar graph of the total weekly amount of physical activity (in hours) and by physical activity it means that the individual has done vigorous exercise or participated sports activities. The graph only shows the hours up to 20, but since there are a lot of extreme outliers, the graph was reduced to give a more pleasant visualization. The vast majority of the sample population is shown to have less than an hour of weekly physical activity. Later on in this report, it can be shown that this graph follows an exponential distribution and its expected value can be found in a range of bootstrapped confidence intervals.

Methods & Models

This section of the report will introduce a variety statistical models that will be used to analyze the variables of the data set and hopefully conclude on the hypotheses in this report. The types of models in this section are a frequentist, Bayesian, and a linear regression model.

(Model 1) Measuring the change in the average BMI per year using linear regression

The first model used here will be a linear regression model. Linear regression models are used to estimate and predict positive or negative linear changes. In reference to Graph 1, an upward linear trend can be observed and thus a this is suitable to use. The model for linear regression can be described as

$$y_i = \alpha + \beta x_i + \epsilon_i$$

where x_i is the ith year and y_i is the average BMI of a Canadian for the ith year. The α represents "y-intercept" and β_1 is the rate of change or "slope" of the line. Finally, ϵ_i represents the error term that varies but at constant amount. The parameter of interest to estimate in this model is β.

- Model Assumptions:
  - there is a linear relationship between x_i and y_i
  - the variables x_i and y_i are independent of each other
  - the ϵ_i's are independent
  - the error term is normally distributed by $\epsilon_i \sim N(0, \sigma^2)$
    - in other words, the ϵ_i's have equal spread or constant variance

The hypothesis this model will try to test is the change in average annual BMI. The null hypothesis in this case would be β=0, or more specifically there is generally no change in annual obesity and the rate of change is 0.

(Model 2) Determining the expected value of weekly physical activity time using Maximum Likelihood Estimation

This next model is a frequentist model as it uses the method of Maximum Likelihood Estimation (MLE) to estimate a parameter of a distribution function. As shown in Graph 3, the bar graph is heavily skewed to the right and the general shape of the graph resembles that of an exponential function, and thus it will be modeled after an exponential distribution. The model for the exponential distribution is

$$Y_i \sim Exp(\theta)$$

where Y_i is the total amount in hours of weekly physical activity for each person i and θ is the rate of that amount. The goal is to estimate the rate parameter θ through MLE. For an exponential distribution, it can be derived that the maximum likelihood estimate for θ is $\frac{1}{\bar{X}}$ or simply the reciprocal of the sample mean. The full derivation of this can be found in the Appendix section. Finally, a 95% bootstrapped confidence interval will be done to give a range of values where it is considered confident where the parameter θ may be. Since the expected value of $Exp(\theta)$ is also known, the 95% confidence interval can also be constructed for the mean of this distribution.

- Model Assumptions:
  - later intervals in time are not affected by previous times and are independent of each other
  - the parameter θ is constant and does not follow a prior distribution like in a Bayesian model

The hypothesis this model will try to test is if there is any statistical significance between overweight (BMI >= 25.0) and non-overweight (BMI < 25.0) people in their weekly physical activity time. The null hypothesis in this case would be to assume that there is no great difference between healthy and non-healthy eaters and their means are essentially the same

(Model 3) Determining the expected BMI value via Bayesian modeling

Graph 2 shows a histogram of the BMIs of all the individuals in the survey. Since the shape of this graph closely resembles a normal curve it will be modeled after a normal distribution under a Bayesian framework. The likelihood and prior for this model is

$$X_1, X_2, ..., X_n \sim N(\mu, \sigma^2), \mu \sim N(27, 1)$$

The X_i's represent an individuals BMI that follow a likelihood function of a normal distribution with a mean of μ and standard deviation of σ. In Graph 2 it shows that the mean could be anywhere between 26-28 and so the prior is chosen to follow a normal distribution with a mean of 27 and standard deviation of 1. Given the likelihood and the prior, the posterior distribution also follows a normal distribution and it derives to

$$\mu|x_1, ..., x_n, \sigma_0^2 \sim N(\frac{\mu_0 \tau_0^{-2} + n\bar{x}\sigma_0^{-2}}{\tau_0^{-2} + n\sigma_0^{-2}}, \frac{1}{\tau_0^{-2} + n\sigma_0^{-2}})$$

and therefore the Bayesian point estimate evaluates to

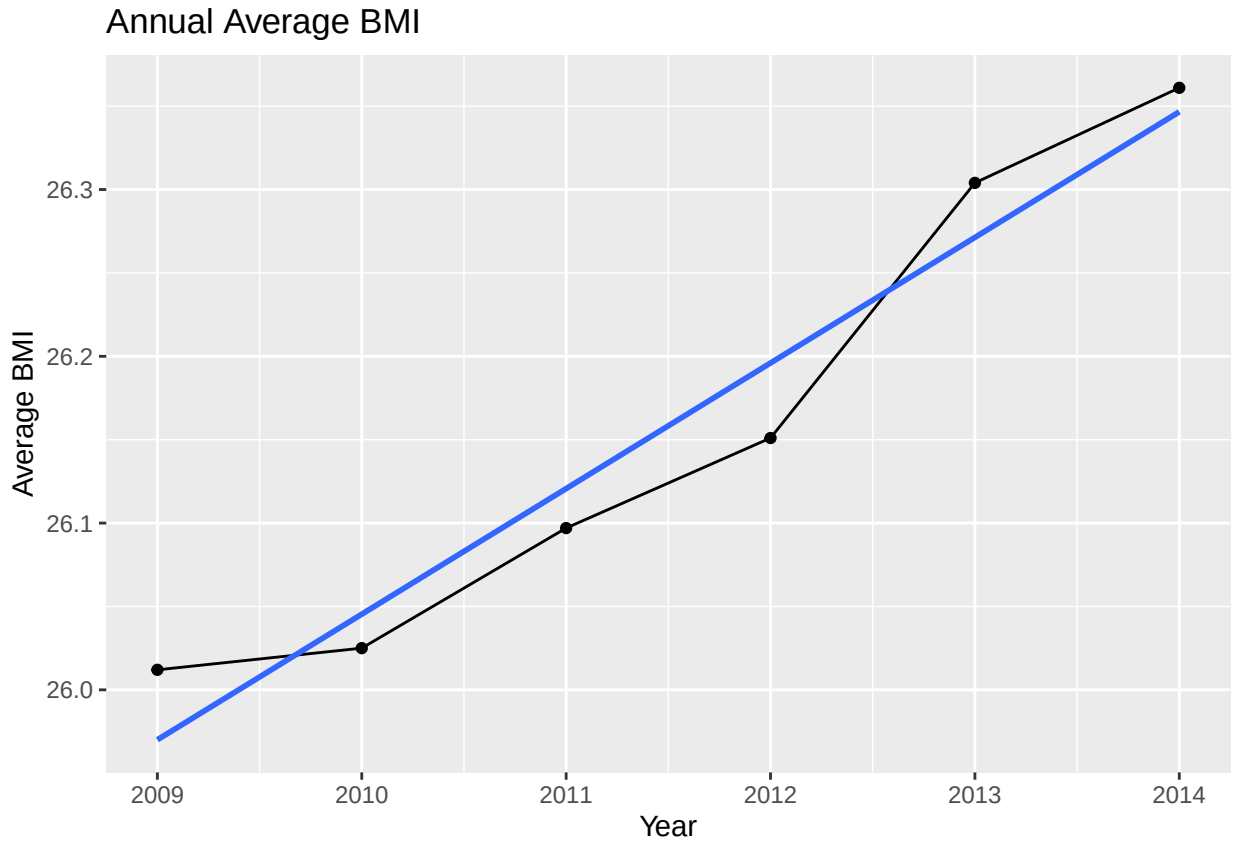
$$\frac{\mu_0 \tau_0^{-2} + n\bar{x}\sigma_0^{-2}}{\tau_0^{-2} + n\sigma_0^{-2}}$$

where μ_0 and τ_0 represents the prior mean and standard deviation (in this case its 27 and 1), n represents the sample population size, $\bar{x}$ represents the sample mean, and σ_0^2 represents the true population variance. A major assumption in this model is it uses the sample variance to substitute for the true population variance. The Bayesian point estimate can be used to determine the mean of the Bayesian posterior and therefore the mean of this Bayesian model. Finally, a 95% credible interval will be given to show the range of values where this point estimate is most likely situated on. The hypothesis this model tests is the statistical significance between healthy (consumes 5 or more fruits/vegetables per day) and non-healthy (consumes less than 5 fruits/vegetables per day). The null hypothesis in this case would be to assume that there is no great difference in average BMI and their means are the same.

Results & Findings

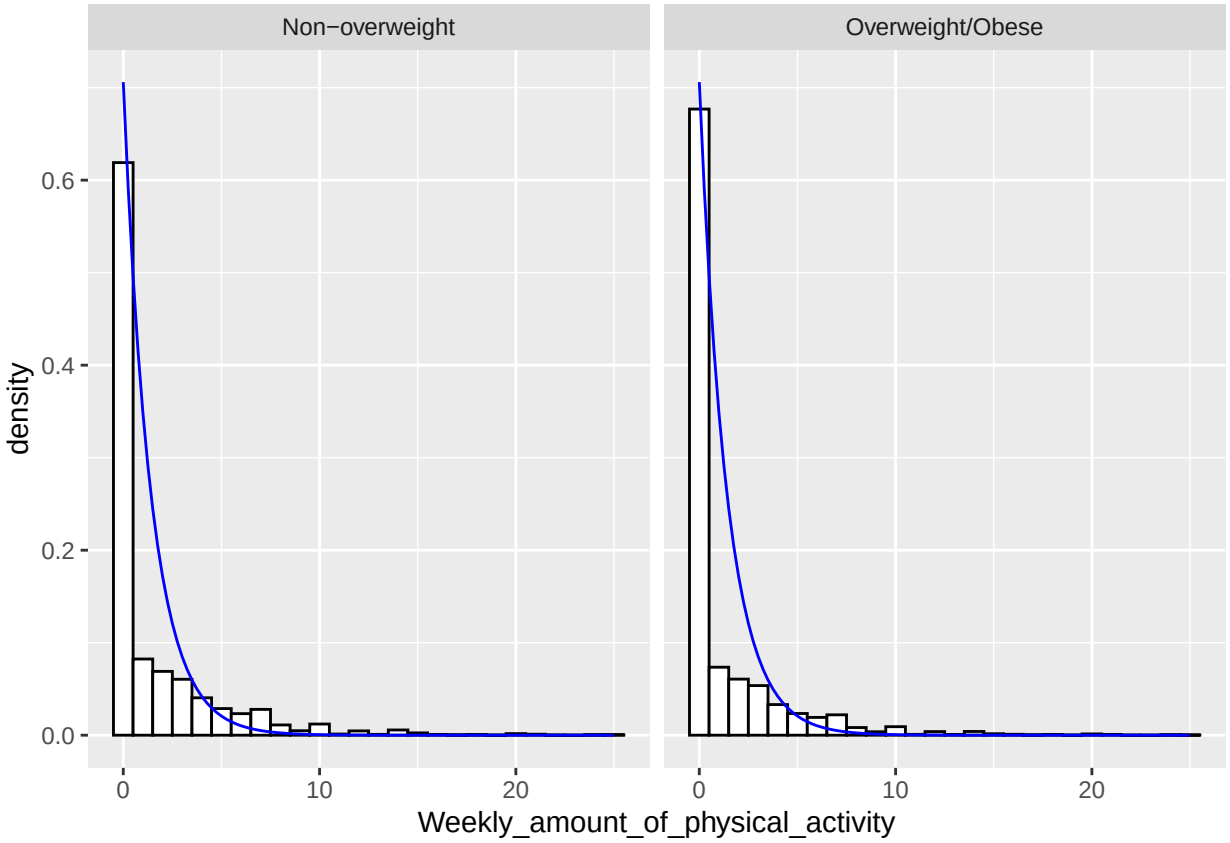
Graph 4 below is the result of the linear regression model and it shows what can be best described as "the line of best fit" for the linear upward trend occurring. By using the lm() function in R, the slope of this line and the estimated β value is determined to be around $\beta = 0.08$. Therefore the rate of change for average BMI per year is about a 0.08 between the years 2009-2014. The lm() function also revealed that the p-value for the two sided t-test is 0.00126 which is less than 0.05, the significance threshold. Therefore the null hypothesis can certainly be rejected and it can be concluded that there is definitely significant change in BMI every year.

Graph 4



Graph 5 below shows the two exponential models between non-overweight and overweight/obese. The MLE θ values were determined to be around 0.62 and 0.76 respectively, and their expected values or means are around 1.62 and 1.32 respectively. Therefore on average under this exponential model, non-overweight individuals spend around 0.3 hours more on physical activity than overweight individuals. The hypothesis test for this model was determined using the t.test() function in R and it resulted in a very minuscule p-value of 2.2e-16, which is significantly less than 0.05. Therefore the null hypothesis can be rejected and this result is deemed very statistically significant. The table below also shows a confidence interval for the means. In both cases, there is a very tight range, meaning its accuracy of where the mean is very high.

(Graph 5)

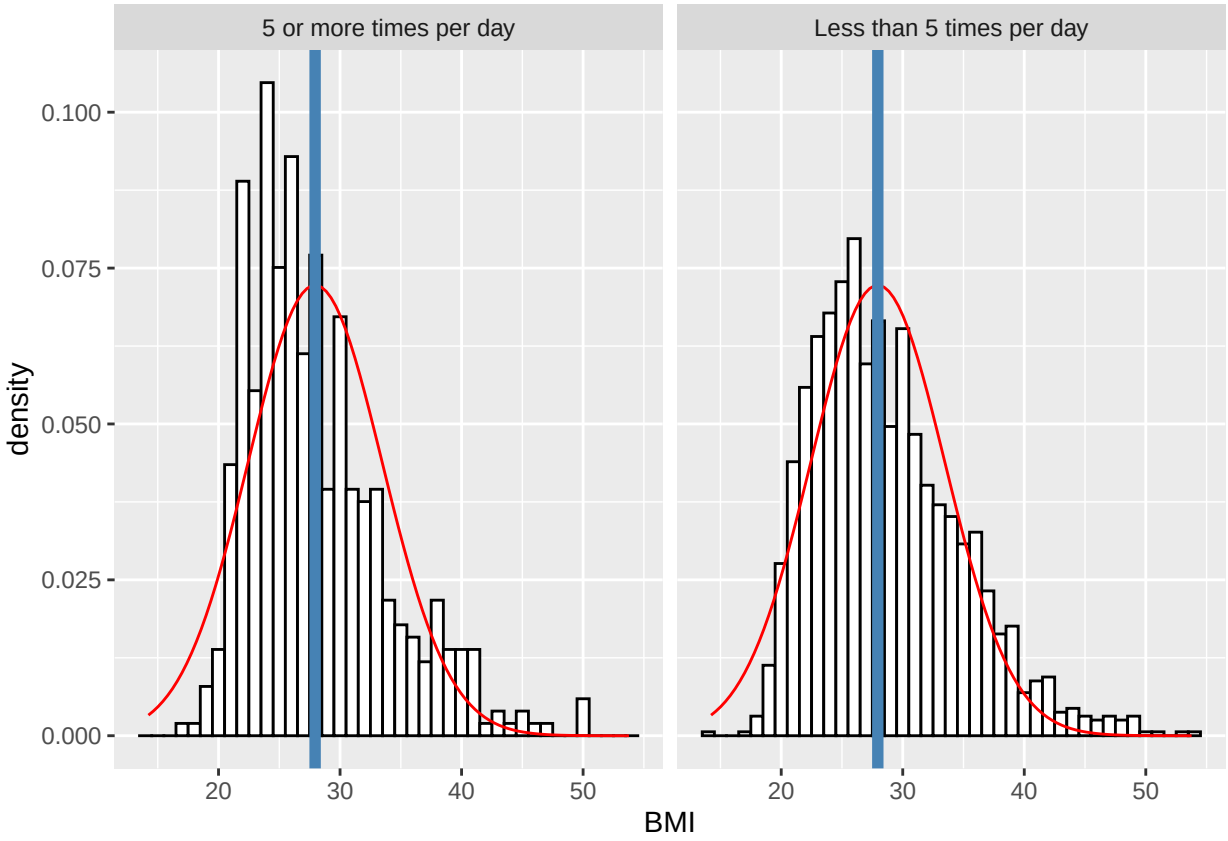


(Table 2) Confidence intervals for weekly physical activity (in hours)

	97.5%	2.5%
Non-overweight	1.615611	1.634515
Overweight/Obese	1.313475	1.329664

Finally Graph 6 shows the Bayesian normal model between healthy and non-healthy eaters. The Bayesian point estimate means were determined to be around 27.94 and 28.65 respectively. The vertical blue line in the graph represents the Bayesian point estimate for all people and the red normal curve represents the Bayesian normal posterior distribution. The hypothesis test is determined by the t.test() function in R, and the p-value was determined to be 8.909e-07 which is significantly less than 0.05. Therefore the null hypothesis is rejected once again in favour of healthy eaters having a lower BMI than non-healthy individuals as a significant statistic. Finally in the table below, the credible intervals are given for both categories and they have a very large interval range with healthy eaters have a slightly narrower interval. The large intervals can be attributed to the large variance of the data. The credible intervals also conclude that the Bayesian point estimate mean has a 95% chance that it lies in those ranges.

Graph 6



(Table 3) Credible intervals for average BMI

Eats heathy less than 5 times a day	16.96969	40.32898
Eats healthy 5 times or more a day	17.13344	38.73970

Conclusion & Summary

Some limitations this report had was mainly due to the data. For example, the linear regression model was only based on 5 years 2009-2014 because the survey sometimes did not ask for the weight and BMI every year. Another drawback was that generally most of the data in the survey was almost all categorical making it was quite difficult to find any continuous data relating to weight and BMI. The main goal this report was determine what possible factors of obesity have statistical significance, but not every factor was included in this report. For future data analysis, there are many other obesity factors that could have been explored such as diseases, family history, mental health, but also in way that is similar to this report such as investigating the frequency of consumption, or the frequency of daily sunlight etc.

In summary the null hypotheses in this report assumed that there were generally no changes or differences in their parameters. Particularly in the linear regression model, it was the assumed that the rate of change was 0. For the frequentist exponential model, the null hypothesis was to assume that there was no difference between the θ's, and for the Bayesian normal model the null was to assume that there was no difference

between the μ's. The results for all hypothesis tests resulted in p-values that were significantly less than 0.05, which is the critical value threshold in determining statistical significance. Since all values were indeed less than 0.05, all results are deemed statistically significant. The linear regression model has shown that there was a increase of 0.08 in average BMI per year during 2009-2014. The frequentist exponential model showed that individuals with non-overweight BMI typically have a higher average physical activity time (1.62 hours) than overweight individuals (1.32) and the Bayesian normal model showed that healthy eaters have a lower BMI on average (27.94) than less healthy eaters (28.65). An interpretation of these results is a quite a simple one, if a person gets involved healthy eating and physical exercise it will help cut down fat, which lowers their weight and therefore lowering their overall BMI.

Appendix

Maximum Likelihood Estimate Derivation

In this model, it is assumed that all x values are greater than or equal to 0. In relation to the data it makes sense since the the value in hours is always positive. The parameter of interest to derive here is θ. The sample data values follow an exponential distribution

$$x_1, ..., x_n \sim Exp(\theta)$$

The exponential distribution function is

$$f(x) = \theta e^{-\theta x}$$

The likelihood function is

$$L(\theta) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \theta e^{-\theta x} = \theta^n e^{-\theta \sum_{i=1}^{n} x_i}$$

The log-likelihood function is

$$l(\theta) = log(\theta^n e^{-\theta \sum_{i=1}^{n} x_i}) = nlog(\theta) - \theta \sum_{i=1}^{n} x_i$$

Derivation:

$$l'(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n} x_i \Rightarrow l'(\theta) = 0 \Rightarrow \frac{n}{\theta} - \sum_{i=1}^{n} x_i = 0 \Rightarrow \frac{n}{\theta} = \sum_{i=1}^{n} x_i \Rightarrow \theta = \frac{n}{\sum_{i=1}^{n} x_i} \Rightarrow \theta = \frac{1}{\frac{\sum_{i=1}^{n} x_i}{n}} \Rightarrow \theta = \frac{1}{\bar{x}}$$

Therefore the MLE for θ is simply just the reciprocal of the sample mean $\bar{x}$.

The second derivative test:

$$l''(\theta) = -\frac{n}{\theta^2} < 0$$

Since the second derivative is always negative, the value derived $\theta = \frac{1}{\bar{x}}$ is indeed a global maximum.

References

1. Uoftlibraries. (n.d.). University of Toronto Libraries. Canadian Community Health Survey, 2017-2018. http://odesi2.scholarsportal.info.myaccess.library.utoronto.ca/webview/index.jsp?object=http%3A%2F%2F142.150.190.128%3A80%2Fobj%2FfStudy%2Fcchs-82M0013-E-2017-2018-Annual-component&gs=7&v=2&mode=download.

2. T.Test: Student's t-test. t.test(). (n.d.). https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/t.test.