Statistics Canada Domestic Travel Survey Research Project

Cho Hsun (Kevin) Yang - 1002210289

23/08/2021

Abstract

The goal of this document is to present the 2004 Canadian Travel Survey dataset and the 2 research questions of determining the domestic travel trends and determining the proportion of trips that use automobiles as their main source of transportation. This document also explains how answering these 2 questions would be applicable in solving global issues. The dataset was mainly collected by Statistics Canada, and we applied additional cleaning and wrangling to create 2 additional datasets to answer the research questions. After the 2 datasets were created, a linear regression model was used to answer the 1st research question, while Maximum Likelihood Estimation and Bayesian Inference were used to answer the 2nd question. Confidence intervals and credible intervals were constructed to give a range of values that the parameter of interest could fall in. Lastly, we conducted a hypothesis test to determine if the proportion of automobile users have decreased in 2004 compared to 1994. For the 1st question, a general increasing trend of 520 trips was observed between January and August. For the 2nd question, it was found that 85.8% of domestic trips use automobiles as their main source of transportation. The results of the hypothesis test concluded that 2004 proportion was indeed less than that of 1994. Even though this analysis answered the research questions, the age of the dataset along with the infrastructure, technological, social, and economic change between 2004 and today in 2021 may limit the applicability of the results. To supplement the results of this analysis, one can consider comparing the trends of domestic trips with international trips. One can also research into the reasons why automobiles are more favorable compared to other transportation modes.

Introduction

Travelling is a big component of today's world, whether it be for business or pleasure. Different modes of transportation has different levels of comfort, speed and environmental impact. Automobiles are seen as the most polluting mode of transportation since each private vehicle carries so little passengers compared to other modes of transport.

This document will introduce the 2004 Canadian Travel Survey Trip File dataset, which was developed to measure the volume, characteristics and economic impact of domestic travel. A more in-depth explanation of the parameters relevant to our analysis will be found in the **Data** section.

This analysis aims to apply statistical techniques on the dataset to answer the questions of:

"What are the trends of domestic trips recorded throughout the year?"

"What proportion of domestic trips use automobiles as opposed to other modes of transportation?"

This analysis will also aim to:

"test the Hypothesis that the proportion of automobile users are similar to 1994 or whether automobile users have decreased in 2004 compared to 1994".

More information on Hypothesis Testing and other statistical methods can be found in the **Methods** and **Appendix** section.

These answers are important to give us an idea of when air pollution from transportation is most heavy throughout the year. The proportion of automobile use can be used as a baseline for future years to compare if infrastructure and technological changes have incentive people to use less polluting public transportation such as subways, streetcars, buses, or bicycles for travel.

The type of analysis and trip data in this document is relevant globally to understand the habits of travelers, allowing governments to understand whether the new national rail is drawing enough customers from driving to decrease congestion, whether investing in an amusement park in this province would likely be more profitable than investing in a museum, or whether hotels should lower their prices in the winter or summer to entice travelers during off season. These are the types of questions that can be answered through similar trip analysis.

Data

Data Collection

This dataset is found **here** on the ODESI(Ontario Data Documentation, Extraction Service and Infrastructure) website, a digital repository for social science data, including polling data. The Canadian Travel Survey, 2004: Trip File used in this document is primarily prepared by Statistics Canada and the support of the Canadian Tourism Commission (CTC) and ten provincial governments. The data is collected through Computer-Assisted Interviewing, where respondents answer questions on devices.

Important Variables

This analysis will only use a portion of the 70347 observations and 106 variables to answer the research questions.

Explanation of the important variables in this dataset:

Year: Year of the observation, it will be 2004 as default for all observations.

Month: Month of the recorded trip observation.

Sum Trips: Aggregated monthly trip observations.

Date: Assigned a date to Sum_trips for plotting graphs.

Mode: Main mode of transportation. Numerical value ranging from 0~6 meaning NA/refused/not stated, Automobiles, Air, Bus, Rail, Boat, bike/motorbike.

Automobile User: Boolean ranging from 0~1(False/True) if the trip mainly relied on automobile for travel. Converted from the Mode variable.

Cleaning and Wrangling

Cleaning and wrangling the data is an important process for formatting and removing errors to create a dataset ready for production and analysis.

This process was done using **R programming** and the **visdat** and **viridis** libraries to check for missing values in the dataset. After parsing through the dataset, no missing values were found.

After checking for missing data, 2 additional datasets was created to answer each of the research questions. The 1st dataset was created with the trip observations aggregated into monthly values to determine the increase or decrease in domestic trips throughout the year. The 2nd dataset contains the Automobile User parameter used to determine the proportion of automobile users. **tidyverse**, **ggplot2**, **qwraps2**, and **knitr** libraries were used to create the datasets.

Dataset 1: Monthly Aggregate Trips

Table 1 below is the summary statistics of the 12 aggregated monthly domestic trips recorded in 2004. Here, the average trips per month is around 5862 trips with a standard deviation of 1128 with the median at 5588 trips and the 1st quartile and 3rd quartile of 5,161.25 and 6,173 trips. Based on the 1.5 Interquartile range rule, all values outside the range of (4070.375, 7105.625) are considered outliers. Looking at the min and max value, we find that there are no low outliers while there are some higher outliers above the upper range of 7105.625. The fact that the mean is larger than the median signifies that the distribution is positively skewed, where most of the data points have smaller value and are closer to the left with fewer larger data points to the right.

Table 1: Summary Statistics of Monthly Trips

	domestic_df (N = 12)
Monthly Trips	
mean (sd)	5,862.25 ± 1,127.67
median (Q1, Q3)	5,588.00 (5,161.25, 6,173.00)
min	4457
max	8056

Figure 1 below is the time-series scatter plot of the monthly aggregated trips throughout 2004. There seems to be 2 main trends, the first being the increase in trips throughout January to August and the second being the decrease in trips throughout August to November. The outlier of a sharp increase in December could be due to winter break for students or the celebration of Christmas, where more people travel back home.
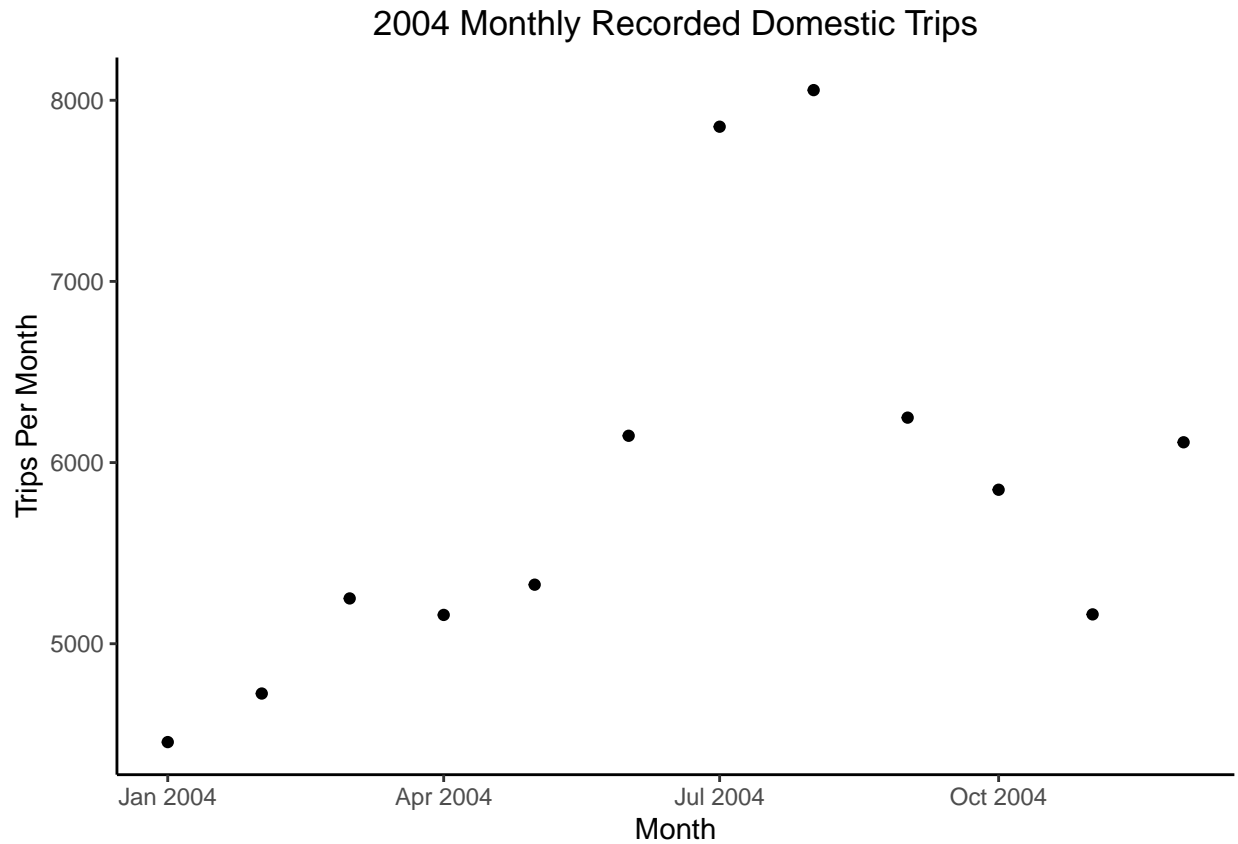

Figure 1: 2004 Monthly Domestic Trips in Canada

Dataset 2: Automobile Users Vs. Other Modes

With dataset 2, we will look at the proportion of trips using automobiles. In a total of 70347 recorded trips, 60355 used automobiles as their main source of transportation during the trip. This means that around 85.8% of all trips use automobiles as their main source of transportation. Table 2 below is a summary table and Figure 2 is a bar plot showing the distribution of trips across the different modes of transportation.

Table 2: Summary Table of Transportation Modes

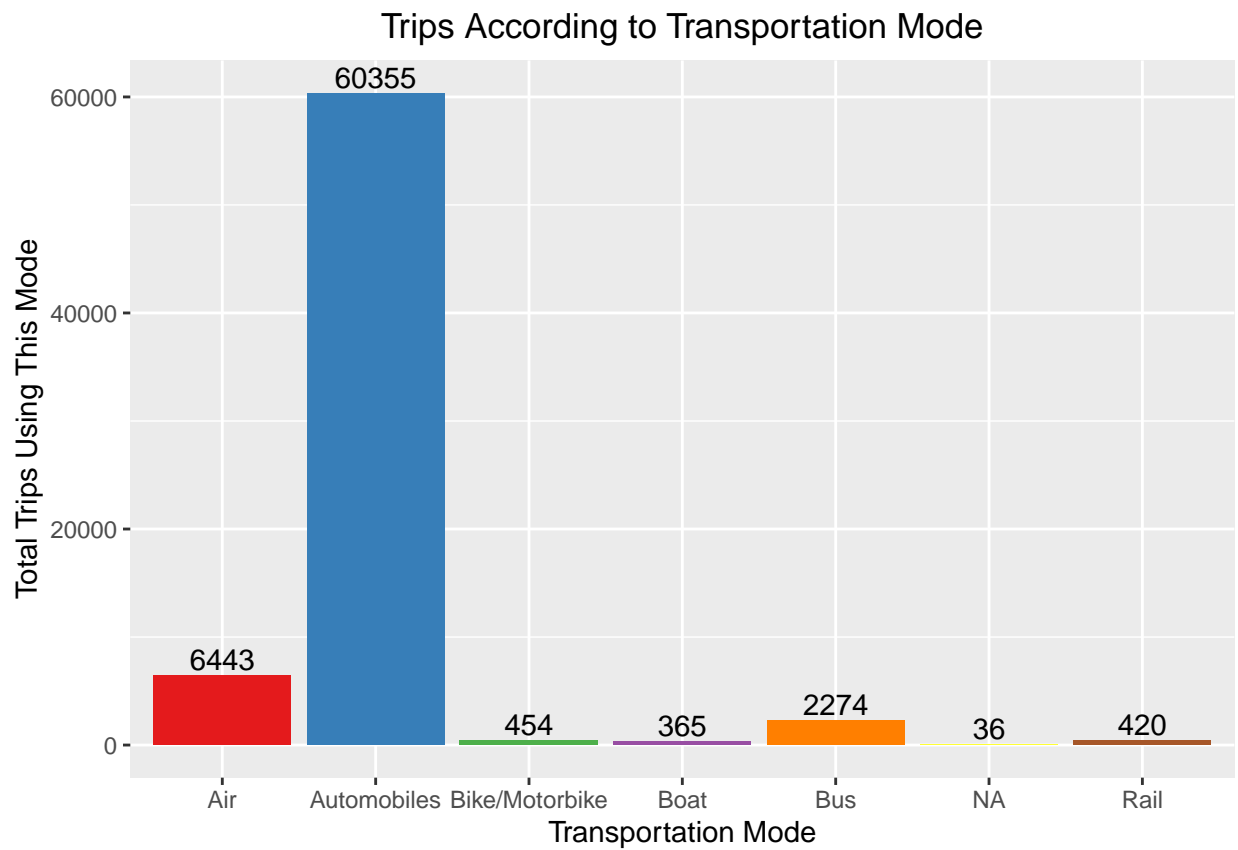| Sum	NA	Automobiles	Air	Bus	Rail	Boat	Bike/Motorbike
70347	36	60355	6443	2274	420	365	454

Figure 2: Bar Plot of Trips According to Transportation Mode

Methods

This section will introduce the different statistical methods that will be used to analyze the data and answer the 2 research questions.

Linear Regression Dataset 1

Since the relationship between the months and aggregated trips seem to have a positive linear trend from January to August, we can use the **Linear Regression Model** to model this trend. We will not model the negative linear trend from August to November, since there are too few data points to make the negative linear regression meaningful.

Below is the Equation for a Linear Regression Model:

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

Y_i**:** The output of this model, which is the number of estimated monthly trips.

β_0**:** The intercept value when x_i is 0.

β_1**:** The slope of the Linear Regression Model representing the rise over run. The vertical difference divided by the horizontal difference is the slope. This is the parameter of interest, since it would signify the average rate of increase or decrease in domestic trips per month.

x_i**:** The independent variable, which represents the months of 2004.

ϵ_i**:** The error of the model, which is the difference between the model with the true value from the dataset. The error is assumed to follow a normal distribution.

Bernoulli Distribution of Dataset 2

For the second dataset, we will use 2 different approaches to estimate the parameter of interest. The first type of method is the Maximum Likelihood Estimation frequentist approach.

First, we will use the discrete **Bernoulli Distribution** model to approximate our dataset. The Bernoulli Distribution is used to model binary outcomes of trials, such as "yes/no" questions or "heads/tails" coin toss. This is suitable since the second dataset has outcomes of "yes/no" to using automobiles as the main transportation mode in each trip.

Below is the Probability distribution for a Bernoulli Distribution:

$$p^k (1-p)^{1-k}$$

k**:** The number of cases of "success", in the case of our dataset, the number of trips which use automobiles as their main source of transportation.

p**:** Probability of "success", the probability of having a trip that uses automobiles as their main source of transportation. This is also the parameter of interest that we will estimate.

Maximum Likelohood Estimation(MLE)

MLE is used to estimate the parameter(s) of a probability distribution by maximizing a **likelihood function** so that it is most likely to observe the data points in the assumed statistical model.

The MLE will estimate $\hat{p}_{MLE}$ of the Bernoulli Distribution by maximizing its likelihood function so that the transportation modes data points will most likely appear in our model.

Here are the steps of using the maximum likelihood estimation:

1. Define the likelihood function.

2. Apply the natural logarithm to the likelihood function to generate the log-likelihood function

3. Differentiate the log-likelihood function and set the result to 0 to find the estimator and estimate.

4. Check if the estimate is maximized by finding the 2nd derivative of the log-likelihood and checking if the result is negative. If it is, then the estimate is maximized.

The full derivation of the maximum likelihood estimator and estimate can be found in the **Appendix**.

Confidence Interval of Dataset 2

The confidence interval is an estimation method that gives a range for the unknown parameter. The interval has a **confidence level** that gives the confidence in which an estimated interval will contain the true value of the parameter. Confidence intervals provide a range of values that are likely if we were to repeat this survey/sample again which is useful for generalizing results. This document will be deriving the 95% confidence interval for the parameter of interest.

Below is the general format for confidence interval:

$$P(L_n \leq \Theta \leq U_n) = 1 - a$$

L_n The random variable lower bound of the confidence interval.

θ: The parameter of interest. It would be $\hat{p}$ for this analysis. An unknown constant.

U_n: The random variable upper bound of the confidence interval.

a: The alpha value in 100(1-a)% confidence interval. For this analysis, a $= 0.05$ for a 95% confidence interval.

Since the transportation mode dataset follows a Ber(p) distribution with $np \geq 10$ and $n(1-p) \geq 10$, we can apply the Central Limit Theorem to have a sample proportion, and it's approximate sampling distribution below:

$$\hat{p} \sim N(p, \frac{p(1-p)}{n})$$

In order to find the lower and upper bounds of the confidence interval, we would solve for the quadratic equation below:

$$(\frac{X}{n} - p)^2 - (z_{a/2})^2 \frac{p(1-p)}{n} < 0$$

$\hat{p}$ The sample proportion of the parameter of interest. Proportion of automobile users for domestic trips.

p: The population proportion p. The true proportion that we are trying to discover.

n: The total number of trips in the sample.

X: The sum of successes in the Bernoulli Distribution. Number of automobile trips.

$z_{a/2}$: The z-score/standard score is the number of standard deviations by which an observed value or data point is above or below the mean value of what is being observed or measured. The alpha value will be 0.05 for the confidence interval of 95%. The z-score values can be found on a z-score table or calculated using **R Programming**.

Hypothesis Test of Dataset 2

Hypothesis testing is used in assessing claims by using evidence/data and is determined based on how probable it is for our evidence to occur.

Here are the usual steps of using the Hypothesis Testing:

1. Stating the null hypothesis (H_0) and alternative hypothesis (H_A). The null hypothesis is the accepted truth or baseline claim that we are trying to disprove. The alternative hypothesis can be a one-sided or two-sided alternative to the null hypothesis.

2. Construct a test statistic, a function of the random variables in the dataset, to assess the evidence against H_0

3. Find the likelihood of the test statistic under the assumption that the null hypothesis H_0 is true. Finding the **p-value**.

4. If the p-value is not small(larger than a) then the original assumption is true, it is likely for us to have observed what we observed. If the p-value is small(less than a), then it is unlikely to observe what we observed if the null hypothesis was true. This result would be **statistically significant** and provide evidence to disprove the null hypothesis.

H_0: p = 0.9. The proportion of automobile users in 2004 is the same as 1994.

H_A: p < 0.9. The proportion of automobile users in 2004 are less than 1994. This is a one-sided test.

The test statistic for the Bernoulli Distribution is :

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

z: The z-score used to find the probability of observing the outcome if the null hypothesis is true.

$\hat{p}$: The proportion of automobile trips derived from the sample.

p_0: The proportion value of the null hypothesis which is 0.9 is this case.

n: The total number of domestic trips.

Bayesian Inference of Dataset 2

The second approach used to estimate the parameter of interest is using the Bayesian Inference approach. Instead of treating the parameter of interest as a fixed unknown constant, the Bayesian model treats it as a random variable, and determines what values are consistent with the data we have observed.

Here are the steps of Bayesian Inference:

1. Setting a **prior**, a probability distribution based on preconceived beliefs or additional data to help determine the estimate. Denoted as $g(\theta)$

2. Find the **likelihood** of data given the estimate. Denoted as $f(x_1, ..., x_n|\theta)$

3. Define the **marginal distribution**. Denoted as $m(x_1, ..., x_n) = \int_{\theta \in \Theta} g(\theta) \cdot f(x_1, ..., x_n|\theta)d\theta$

4. Find the **posterior distribution** of the parameter of interest. The posterior combines information from the dataset and the prior distribution to estimate the parameter of interest. Denoted as

$$\pi(\theta|x_1, ..., x_n) = \frac{g(\theta) \cdot f(x_1, ..., x_n|\theta)d\theta}{m(x_1, ..., x_n)}$$

Since the model of transportation mode choice is a Bernoulli Distribution, the likelihood is also a Bernoulli Distribution. Looking at the conjugate pairs of Bayesian Statistics, we can use a Beta Distribution prior to obtain a Beta Distribution posterior distribution. Conjugate pairs are used if possible to simplify the calculation of the posterior distribution.

For a $Beta(\alpha, \beta)$ prior and a $Ber(p)$ distribution likelihood, the resulting posterior is:

$\hat{p}_{Bayes}|x_1, ..., x_n \sim Beta(n\bar{x} + \alpha, n(1 - \bar{x}) + \beta)$

After some testing of choosing parameters for the prior, the Beta(8,1.5) distribution was chosen as the prior since the expected value of 0.842 is close to that of the 2004 sample p of 0.858 and the 1994 sample p of 0.9, making it a suitable prior for our dataset. Figure 3 below is a graph showing this distribution.
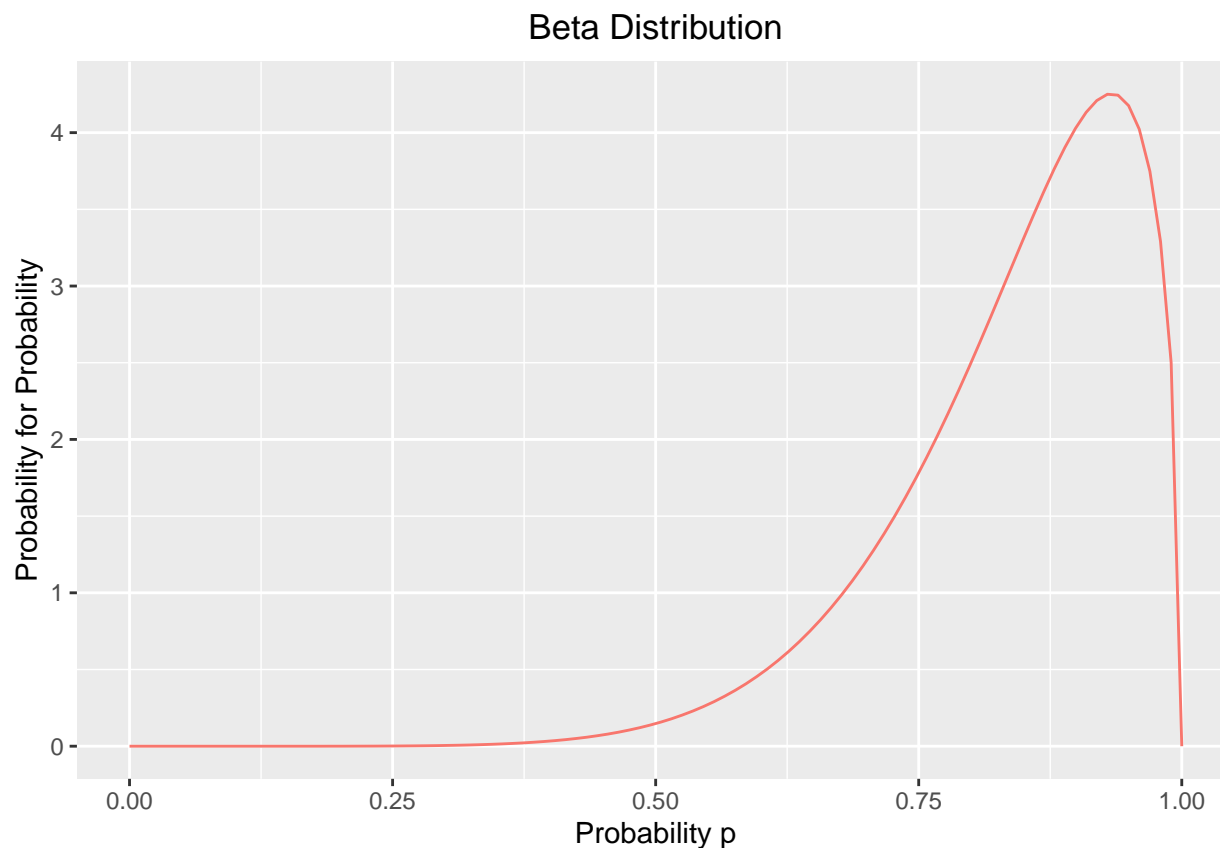

Figure 3: Beta(8,1.5) Prior Distribution

Credible Interval of Dataset 2

Similar to the idea of Confidence Interval, Credible Intervals are created from the posterior distribution of θ and is interpreted as an interval containing the θ with (1-a)100% probability since the Bayesian estimate θ is a random variable instead of a constant.

Below is the general format for credible interval:

$$P(L_n \leq \theta \leq U_n | x_1, ..., x_n) = 1 - a$$

L_n The random variable lower bound of the confidence interval.

θ: The parameter of interest. It would be $\hat{p}$ for this analysis. A random variable.

U_n: The random variable lower bound of the confidence interval.

a: The alpha value in 100(1-a)% confidence interval. For this analysis, a = 0.05 for a 95% credible interval.

Results

Linear Regression Results

After generating a linear regression model in R, below is the rounded results:

$$Y_i = 3530.857 + 520.226x_i$$

The parameter of interest $\hat{\beta}_1$ is 520.226. Throughout the months of January 2004 to August 2004, there is an average monthly increase of 520 domestic trips. The error of each point would be the difference between the true data point and the data point generated in the Linear Regression Model. The resulting slope makes sense since it balances out the slight decrease from March to April with the sharp increase from June to July and the regular increasing values in the other months.

Figure 4 below is a graphical summary of the data with the Linear Regression Model showing the difference between the modeled values and true values, and how the slope balances out the variances in increasing and decreasing slopes throughout the months.
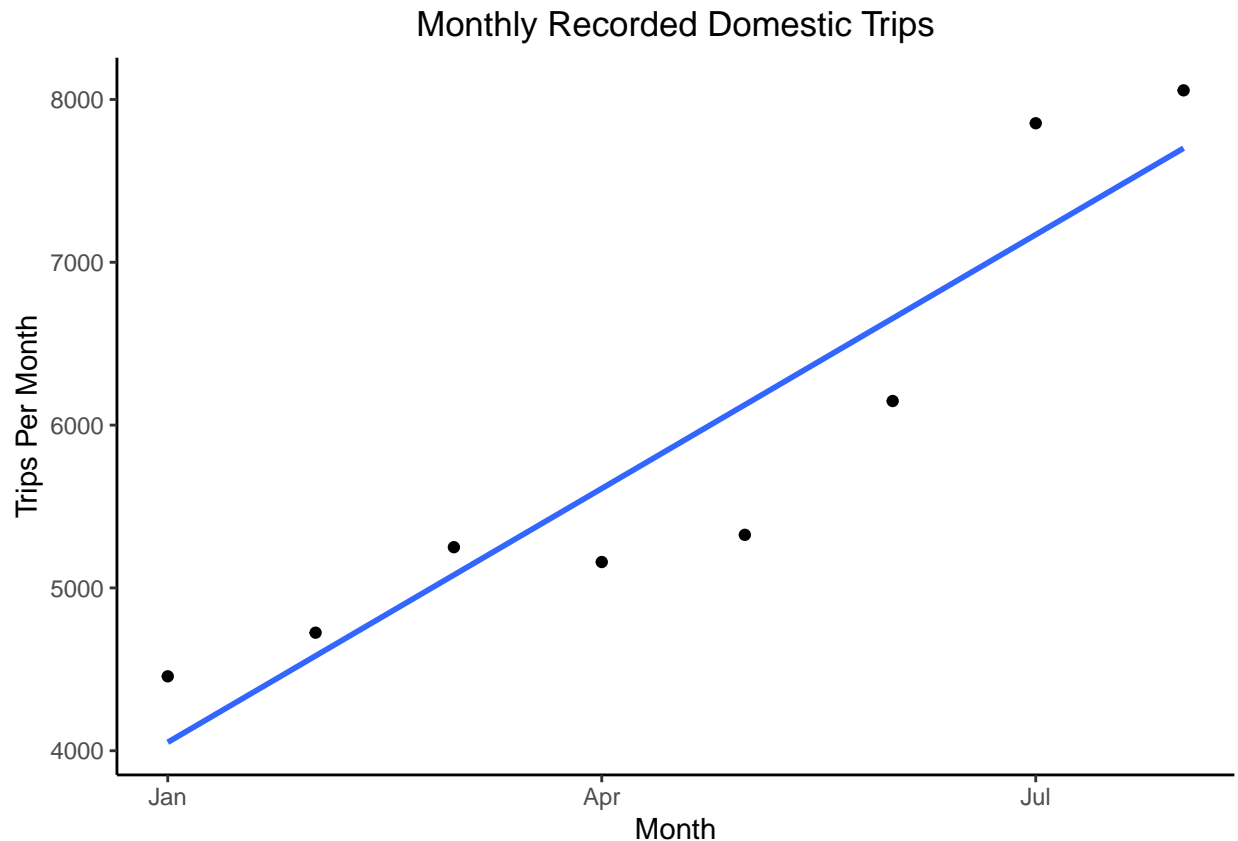

Figure 4: Linear Regression Model fitted onto Dataset

Maximum Likelohood Estimation Results

After applying MLE, the estimator for $\hat{p}_{MLE}$ is the sum of values divided by the number of trips. The estimate is 0.8580:

$$\hat{p}_{MLE} = \frac{\sum_{i=1}^{n} X_i}{n} \approx 0.8580$$

This means that, according to the MLE estimate, 85.8% of domestic trips use automobiles as their main mode of transportation. This result makes sense since it is consistent with the result proportion of automobile users obtained in the **Data** section. The full derivation can be found in the **Appendix** section.

Confidence Interval Results

In order to use the CLT, we must check that the statements of $np \geq 10$ and $n(1-p) \geq 10$ are true.

Subbing in the values of n = 70347 and p = 0.858 from the dataset, the result satisfies the 2 conditions for using CLT:

$$60357.726 \geq 10$$
$$9989.274 \geq 10$$

Now that CLT can be used, we will use the quadratic equation stated in the **Methods** section to solve for the bounds:

$$(\frac{X}{n} - p)^2 - (z_{a/2})^2 \frac{p(1-p)}{n} < 0$$

We can substitute in the numbers of X = 60355, n = 70347, p = 0.858, $z_{a/2} = 1.96$ from the dataset and **R programming** to get the resulting quadratic equation:

$$1.00005p^2 - 1.71597p + 0.73610 < 0$$

Solving for this, the confidence interval is:

$$CI_{95} = (0.85674, 0.85915)$$

We are 95% confident that the interval of (0.85674, 0.85915) contains the true value of the parameter $\hat{p}$, the proportion of domestic trips that use automobiles as their main source of transportation.

Hypothesis Test Results

Here are the null hypothesis and the alternative hypothesis:

H_0: p = 0.9.

H_A: p < 0.9.

Now, let's find the test statistic:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.858 - 0.9}{\sqrt{\frac{0.9(1-0.9)}{70347}}} \approx -37.132$$

Since we are conducting a one-sided test to check if p < 0.9, the p-value would be:

$$p(Z < -37.132) \approx 4.245887e^{-302} \approx 0$$

The resulting value is much smaller than our alpha value of 0.05, so there is very strong evidence against H_0 that the 2004 proportion of automobile users are the same as the proportion of 1994 automobile users. In other words, there is strong evidence that the proportion of 2004 domestic trips with automobiles as their main mode of transport is less than the proportion in 1994. We reject H_0 in favor of H_A. The results seem reasonable, since the proportion has decreased in 2004 compared to 1994.

Bayesian Model and Credible Interval Results

Here are the established likelihood and prior distribution:

Likelihood: $\text{Ber}(\hat{p}_{Bayes})$

Prior: $\text{Beta}(8,1.5)$

Plugging in the values from the dataset, the posterior would be:

$$\hat{p}_{Bayes}|x_1,...,x_n \sim Beta(n\bar{x}+\alpha, n(1-\bar{x})+\beta) = Beta(60363, 9993.5)$$

The next step is to find the Bayesian Estimate using the expected value of the posterior:

$$\hat{p}_{Bayes} = E[X] = \frac{\alpha}{\alpha+\beta} = \frac{60363}{60363+9993.5} \approx 0.8580$$

The Bayesian estimate $\hat{p}_{Bayes}$ is approximately 0.8580, meaning that around 85.8% of domestic trips are recorded as using automobiles as their main mode of transportation. This result makes sense since it is consistent with the MLE estimate along with the sample mean.

In order to find the credible interval, we would use the posterior distribution of $\hat{p}_{Bayes}$ and **R programming** to find the upper and lower bounds of the interval.

The credible interval is:

$$CI_{95} = (0.85537, 0.86053)$$

For the credible interval, there is a 95% probability that p of choosing car over other modes is between 0.85537 and 0.86053 given our data. The credible interval is found to be wider than the confidence interval.

Conclusion

The goal of this document was to answer the 2 questions of:

"What are the trends of domestic trips recorded throughout the year?"

"What proportion of domestic trips use automobiles as opposed to other modes of transportation?"

We used linear regression to model the trend of domestic trips throughout the year, namely the positive trend from January to August, resulting in an average increase of 520 trips per month. To answer the 2nd question, we used the Maximum Likelihood Estimation and Bayesian Inference to estimate the proportion of domestic trips who uses automobiles as their main mode of transportation. They both yielded estimates of 0.8580 or 85.8% of all trips. The 95% confidence interval is between the range of (0.85674, 0.85915) while the 95% credible interval is between (0.85537, 0.86053).

After conducting the **Hypothesis test**-, it was found that there is strong evidence that the proportion of 2004 domestic trips with automobiles as their main mode of transport is less than the proportion in 1994. We reject the null hypothesis and accept the alternative hypothesis.

From these results, we can conclude that the amount of domestic trips increases throughout the year until August, where it would start decreasing. This means that there is a gradual increase in air pollution caused by transportation, with the peak in August. The result of the second research question suggests that a large proportion(85.8%) of domestic travelers prefer using automobiles rather than other modes of transportation. This is concerning since automobiles tend to be the highest polluting mode of transportation due to the high volume and the small number of people each vehicle carry. The hypothesis test does show promising result though, since automobile usage has decreased from 90% in 1994 to 85.8% in 2004 which can signify a decrease in reliance on automobiles, potentially leading to less pollution.

Although this document answers the 2 research questions, there are limitations to this analysis that may not make it as relevant to the problems of today. One of the limitations is how dated the dataset is. The Canadian Travel Survey Trip File is from 2004 while this analysis is conducted in 2021. Due to infrastructure and technological improvements, the analysis would likely have yielded different results with more current data. If the goal was to determine the air pollution caused by automobiles, the analysis would need to account for the recent invention of electric vehicles that could decrease the value by a lot. Another consideration is of the social and economic changes comparing 2004 to today. Covid-19, the economic downturn, and the rise of work from home culture can greatly decrease the total domestic trips in Canada, yielding different trends of domestic trips in 2021 compared to 2004.

Although there are limitations to this analysis, there are many ways to improve and build upon the results. For research question 1, we can also conduct an analysis for international trips as well to see how similar or different the trends are between domestic trips and international trips. If they have similar peaks and troughs, could it be that people travel more during school breaks and less during school time? If they are different, what could be the factor that affects people's travel habits between domestic and international trips? These are the type of questions one could expand on.

With the results of the high proportion of automobile users from the 2nd research question, we can investigate why they choose this transportation mode over the other ones. Is it due to convenience? Cost? Comfort? We can conduct a survey to determine the top reason(s) why people tend to choose automobiles as the main mode of transportation for domestic trips.

Bibliography

This document was completed using **R programming** in the **RStudio** integrated development environment software. Documentation for R can be found **here**. Below is a list of the specific libraries used for this analysis:

tidyverse

skimr

visdat

ggplot2

viridis

qwraps2

knitr

reshape2

The 2004 Canadian Travel Survey Trip File used to conduct the analysis is found **here** at the ODESI website.

Appendix

This section will demonstrate the derivation of the Maximum Likelihood Estimation of the proportion of automobile users.

As stated in earlier sections, we assumed dataset 2 follows the Bernoulli Distribution which is derived as:

$$Ber(p) \sim p^k (1-p)^{1-k}$$

The likelihood is derived as:

$$L(\hat{p}) = p(x_1|\hat{p}) \cdot (x_2|\hat{p}) \cdot ... \cdot (x_n|\hat{p}) = \prod_{i=1}^{n} p(x_i|\hat{p})$$

With the Bernoulli Distribution the Likelihood is:

$$L(\hat{p}) = \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i}$$

Next, we calculate the log likelihood:

$$l(p) = \sum_{i=1}^{n} x_i \ log \ p \ + \sum_{i=1}^{n} (1-x_i) \ log \ p$$

We take the 1st derivative of the log likelihood setting it to 0:

$$\frac{dl(p)}{dp} = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{\sum_{i=1}^{n}(1-x_i)}{1-p} = 0$$

After isolating the parameter of interest, the MLE estimator is:

$$\hat{p}_{MLE} = \frac{\sum_{i=1}^{n} x_i}{n}$$

We can sub in the sum of $x_i = 60355$ and n = 70347 from the dataset to find the MLE estimate:

$$\hat{p}_{MLE} = \frac{60355}{70347} \approx 0.8580$$

We take the 2nd derivative of the log likelihood and check if it is negative:

$$\frac{d^2 l(p)}{dp^2} = \frac{-\sum_{i=1}^{n} x_i}{p^2} - \frac{\sum_{i=1}^{n}(1 - x_i)}{(1 - p)^2}$$

Since $p \; \epsilon \, [0, 1]$ and $x_i \; \epsilon \; 0, 1$, the second derivative is negative, and therefore the estimate is maximized.