Modeling the Year Over Year Change of Employed Persons in Canada Using Logistic Regression

Aditya Khan

March 2021

1. Introduction

The coronavirus pandemic has had an impact on numerous aspects of our lives. In particular, the pandemic has taken its toll on the global economy, including bankruptcy, losing business and a sharp increase in unemployment rate. In a recent poll, one in every three unemployed Canadians reported that their job was lost due to the pandemic ("NEW POLL", 2020). Due to the cultural and economic diversity in Canada, the pandemic might have different impacts in different regions of the country. Thus, my primary objective in this study is to explore the effects of the pandemic on the number of employed persons in Canada taking into account its regional diversity. Particularly, I will use a logistic regression to analyse pre- and post-pandemic employment data, to shed light on the difference in employment before and after the pandemic ("Labour force", 2021).

The pandemic affecting our economic outcomes provides a salient opportunity to explore statistical models with real-world applications. To this end, my secondary objective is to explore how to formulate the logistic regression model under this setup, and how to present and interpret statistical results in the context of my primary objective as mentioned above. In doing so, I will discuss the justification for using a logistic regression model in this context, and I will interpret the results in probabilistic terms using the odds ratio.

I would like to explore this topic because of my keen interest in understanding the current economic situation using statistical modeling. I am particularly interested on statistical modeling and inference along with their applications in different areas of knowledge. Beyond personal interest, the academic rationale for this study is that the impact of the pandemic on various sectors of Canadian economy is not a settled topic, and there is always a scope for further exploration. This motivated me to conduct this study, which involves both statistical modeling and practical understanding of the change of economy over time.

2. Data and Research Questions

Monthly data on the number of employed persons across Canada are available in the Statistics Canada database, which is a government source ("Labour force", 2021). In addition to national aggregate data (which I will refer to as "national data"), there are also data segregated by province. As a result, I can define three distinct regions across Canada to explore regional effects on employment.

- Region 1: Atlantic Canada, which includes Nova Scotia, New Brunswick, Newfoundland and Labrador, and Prince Edward Island.
- Region 2: Central Canada, which includes Ontario and Quebec.
- Region 3: Western Canada, which includes Alberta, Saskatchewan, Manitoba, and British Columbia.

Note that Statistics Canada does not have official employment data for the three Canadian territories for this specific endeavour, and therefore I cannot include them in this study. This is an unfortunate limitation of my study.

Although data are available from 1977 to 2021 ("Labour force", 2021), I need to focus on a relatively short period to highlight the effects of the pandemic. This is because a longer time frame for my data would shift the focus away from the pandemic's impact on employment. At the same time, having too small of a sample size makes the statistical inference unreliable. Therefore, considering this all, I must choose a study period so it specifically reflects the economic change due to the pandemic. Eventually, I decided on considering monthly employment data from January 2018 to January 2021, which would provide enough information to model the change in the number of employed persons over time, without making the sample size too small.

Year over year (YOY) change is an economic comparison that is used in reference to various economic indicators (Cieslak, 2021). In this study, I will use this measure to describe the change of employed persons in Canada in relation to time and region. It is defined as

$$\text{YOY change (\%) for current month} = \frac{\text{Current month} - \text{Reference month}}{\text{Reference month}} \times 100, \quad (1)$$

where current month = number of employed persons within the month in question, and

reference month = number of employed persons within the month 12-months prior.

The rational behind why I prefer this measure over a more common measure such as the change in employment rate are: (a) it is an accepted measure of employment percent change by Statistics Canada ("Labour force", 2021), (b) as employment status can change on a monthly basis (Majaski, 2021), the use of the YOY measure makes sense here, and (c) although employment rate is a more common measure of employment change, YOY change should illustrate greater variation in data points due to the pandemic. This is because YOY change directly compares the number of employed persons, which is subject to greater variability than change in employment rate (it only represents employment as a proportion to the population). This was confirmed when I experimented with different measures before this exploration. Thus, using YOY change would further highlight the abnormality of a pandemic-affected employment situation, which is the point of this paper.

The national data for the YOY measure are included in the Statistics Canada database ("Labour force", 2021). To compute this measure for a region, I first computed the total number of employed persons for that region, and then I used Equation (1). For example, I computed the YOY change for the Central Canada for January 2018 as follows:

$$\text{YOY change for Central Canada for January 2018} = \frac{(\text{ON+ QC})_{\text{Jan 2018}} - (\text{ON+ QC})_{\text{Jan 2017}}}{(\text{ON+ QC})_{\text{Jan 2017}}} \times 100, \quad (2)$$

where ON = number of employed persons in Ontario during the month in the subscript, and

QC = number of employed persons in Quebec during the month in the subscript.

To provide a representative sample of my data, the computation of YOY change for the Central Canada from January 2018 to June 2018 is demonstrated in Table 1 (the complete data are given in Appendix B for regional data). Table 2 summarizes the national and regional YOY change in Canada from January 2018 to June 2018 (see Appendices for complete data). Immediately, we can see that the YOY change for the Atlantic region is consistently below the national change, whereas this change for the Central region is consistently higher than the national change. We also see that the change for the Western Canada is sometimes higher and sometimes lower than the national change. This is the typical trend for YOY change for the whole study period.

Table 1: YOY change in Central Canada from January 2018 to June 2018.

Month	Reference Month's Employed Persons*	Current Month's Employed Persons	YOY Change (%)
January	11,123,700	11,309,500	1.6703
February	11,118,500	11,324,300	1.8510
March	11,127,100	11,375,900	2.2360
April	11,125,500	11,364,400	2.1473
May	11,158,900	11,366,200	1.8577
June	11,179,200	11,404,000	2.0109

* reference month refers to the corresponding month of 2017

Table 2: National and regional YOY change (%) in Canada from January 2018 to June 2018.

Month	National Data	Atlantic Canada	Central Canada	Western Canada
January	1.6688	0.15474	1.6703	1.9489
February	1.7193	1.4474	1.8510	1.5193
March	1.9380	0.61999	2.2360	1.6214
April	1.7550	1.0944	2.1473	1.1442
May	1.4105	1.1055	1.8577	0.61688
June	1.4795	1.2721	2.0109	0.52358

Note that from January 2018 to January 2021, the national mean and median of YOY change are –0.56% and 1.54%, respectively. What this shows me is that the distribution of YOY changes is highly skewed to the left (likely due to the pandemic's disproportionate effect later in my data set's time frame. Now, both mean and median are measures of central tendency. However, since the mean is affected by the extreme values of a skewed distribution, the median has to be used for the measure of central tendency for skewed distributions. So, I will consider the median instead of the mean as the threshold to define the stability/instability of the employment situation. Using the median, I can define the variable Y in a binary manner:

$$Y = \begin{cases} 1 \text{ if YOY change} \leq 1.5\% \\ 0 \text{ if YOY change} > 1.5\% \end{cases}. \quad (3)$$

With this definition, I will consider $Y = 0$ to indicate **"stability in employment"**, and $Y = 1$ to indicate an **"inferior condition"**. I will use this definition for both national and regional data to keep one reference point for all analyses. Given all of this, I had four separate questions immediately come to mind. I would like to answer them in this study.

a) Is the risk of an inferior condition increasing over time?

b) Does the risk of an inferior condition significantly differ by region?

c) Is the risk of an inferior condition higher/lower for one region compared to another region?

d) What is the overall effect of time on YOY change?

Note that the study period ranges from a pre-pandemic time point to a post-pandemic time point. Thus, if I find that the risk of an inferior condition increases over time (Question (a)), this will indicate a negative impact of the pandemic on the stability in employment. To address these questions, I will formulate the logistic regression model as described in the following section.

3. Method

A regression model is a natural choice to understand the effects of time and region on employment condition (Kleinbaum & Klein, 2010). Since I am considering the response variable Y as binary, a logistic regression is an appropriate model to use (Kleinbaum & Klein, 2010). Within this study, I have data for 37 months (January 2018 to January 2021). This means that a change in time may influence whether the risk of an inferior condition increases. So, I can define the time variable as $x_1 = 1, 2, \dots, 37$. That is, each month in the study has a numerical value assigned to it. For analysis of the regional data as time goes on, I need to consider an identifier for different regions in Canada. On the other hand, if I want to understand the overall effect of time, I have to analyse the national data with only one predictor variable x_1, separate from the three regions. In the following section, I will present the formulation of the logistic regression model for the regional data, and for national data.

3.1 Formulation of the Logistic Regression Model

In Section 2, I defined three regions across Canada. Now, I can define two dummy variables to identify those three regions, as follows:

$$x_2 = \begin{cases} 1 \text{ if Central Canada} \\ 0 \text{ otherwise} \end{cases} \tag{4}$$

$$x_3 = \begin{cases} 1 \text{ Western Canada} \\ 0 \text{ otherwise} \end{cases} \tag{5}$$

Using Central Canada as an example, x_2 takes the value of 1, when considering that region, and 0 otherwise. With these definitions, the I can now call the Atlantic region, the reference category. What that means is that when $x_2 = x_3 = 0$, that must define the Atlantic region. The reason I chose not to define another separate variable for the Atlantic region, is that it would be redundant, as

these variables already leave open a scenario where the Atlantic region is represented. Above all, the benefit of using dummy variables instead of creating different regression models for each region, is that I can directly compare each region in one combined regression.

The logistic regression model is expressed in terms of the probability of the occurrence of an event. Recall that the response variable is defined as $Y = 0$ for a stable employment condition and $Y = 1$ for an inferior condition. The probability of $Y = 1$ given the explanatory variables x_1, x_2 and x_3, denoted by $P(x)$, can be expressed as follows:

$$P(Y = 1) = P(x) = \frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3}}, \tag{6}$$

where β_0, β_1, β_2 and β_3 are called regression coefficients, which are unknown. For clarity's sake, $[\beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3]$ in equation (6) is the power, with e being the base. Those regression coefficients will be estimated based on my sample data (see Section 3.2 for the estimation method and Section 3.3 for interpretation of the regression coefficients). Note that

$$P(Y = 0) = 1 - P(x) = \frac{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3}} - \frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3}}. \tag{7}$$

Thus,

$$P(Y = 0) = \frac{1}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3}}. \tag{8}$$

After this, we tend to do one more transformation in statistics. I can do this by taking the natural logarithm of the ratio comparing $P(x)$ against $1 - P(x)$. Then, the transformation

$$g(x) = \ln\left[\frac{P(x)}{1-P(x)}\right] = \frac{\frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3}}}{\frac{1}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3}}} = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 \tag{9}$$

is called the logit transformation, denoted by $g(x)$. The reason why I did this transformation is that it shares some useful properties with a linear regression model. For example, it is linear in its

parameters, it is continuous, the range goes from $-\infty$ to ∞ (Hosmer & Lemeshow, 2000). The other benefit I can see is that $g(x)$ is essentially the natural logarithm of an odds. Odds compare the probability of a favourable outcome against that of an unfavourable outcome, written as $\frac{P(x)}{1-P(x)}$ ("What is an Odds", 2020).

The model (6) takes into account both time (x_1) and region (x_2 and x_3) to describe the probability of $Y = 1$, and will be used to describe the regional data. For the national data, I have only one predictor variable x_1, as I am doing this regression separately. Thus, the logistic model has the simple form

$$P(Y = 1) = P(x) = \frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_1 x_1}}. \tag{10}$$

Beyond the fact that I wanted to explore both simple and multiple regressions, the benefit to doing a simple regression with national data separately, is that I can graphically view the national data independent of the effect of regional data.

3.2 Estimation

As the regression coefficients are unknown quantities, I have to estimate them with my data. To do this, I can use the maximum likelihood method for estimation, which is effectively based on maximising a function called the likelihood function (Hosmer & Lemeshow, 2000). The likelihood function expresses the probability of the observed data as a function of the unknown parameters (the regression coefficients for our model). What this can tell me is how likely the observed data is, as a function of the parameter values (Hosmer & Lemeshow, 2000). Thus, maximising the likelihood estimates the parameters such the parameters most accurately represent my data set (Balaban, 2018). Duly, the maximised parameter values are called maximum likelihood estimates.

Remember that I defined the response variable Y for the logistic regression model so that it can only take two possible values: 0 and 1. In the likelihood function, I can represent this as $P(Y = 1) = P(x)$ when $Y = 1$, and $P(Y = 0) = 1 - P(x)$ when $Y = 0$. When I have a given data point i, I can combine these two probabilities into one likelihood expression as shown below:

$$L_i = [P(x_i)]^{y_i}[1 - P(x_i)]^{1-y_i}. \quad (11)$$

Note that (11) reduces to $P(x_i)$ when $y_i = 1$ and $1 - P(x_i)$ when $y_i = 0$, The result of this is that the i^{th} one of my data points is contributed to the likelihood function (Hosmer & Lemeshow, 2000). I then can obtain the full likelihood function as the product of the terms (11) over all data points by doing the following:

$$L = L_1 L_2 \dots L_n = \prod_{i=1}^{n} L_i = \prod_{i=1}^{n} [P(x_i)]^{y_i}[1 - P(x_i)]^{1-y_i}, \quad (12)$$

where $\Pi_{i=1}^{n}$ is the product notation, representing the product of n terms, and $n = 37$ (the sample size; there are 37 months in this study) for my data.

However, in the field of statistics, the usual convention is to take the log of the likelihood, because it turns the products into sums, which are easier to deal with (Hosmer & Lemeshow, 2000). Furthermore, the parameter values that maximise the likelihood function also maximise the log likelihood function, because the logarithmic function is a monotonic function (Hosmer & Lemeshow, 2000). Thus, maximising the likelihood function is effectively equivalent to maximising the log likelihood function. From (12), I can write the log likelihood function as

$$\ln L = \sum_{i=1}^{n} [y_i \ln P(x_i) + (1 - y_i) \ln(1 - P(x_i))]. \quad (13)$$

Now that I have to find the values of the parameters that maximise $\ln L$, I have to differentiate $\ln L$ with respect to the parameters. In doing so, I must set the resulting equations equal to zero (Hosmer

& Lemeshow, 2000). By solving these equations, I can achieve the maximised estimates of the parameters.

As an example, consider the model (6) for the regional data. The log-likelihood function can be written as

$$\ln L = \sum_{i=1}^{n} \left\{ y_i \ln \left[\frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3}} \right] + (1-y_i) \ln \left[\frac{1}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3}} \right] \right\}. \quad (14)$$

Then, solving the set of equations $\frac{\partial \ln L}{\partial \beta_0} = 0, \frac{\partial \ln L}{\partial \beta_1} = 0, \frac{\partial \ln L}{\partial \beta_2} = 0, \frac{\partial \ln L}{\partial \beta_3} = 0$ gives me the maximum likelihood estimates of β_0, β_1, β_2 and β_3. Note that ∂ is used to denote a partial derivative. With that said, the problem is that these are nonlinear equations and cannot be solved analytically (Hosmer & Lemeshow, 2000). Therefore, I must use a numerical method (which is effectively guessing and checking) for maximisation. I did this with the statistical program R, to find the maximum likelihood estimates of the regression coefficients. I chose to learn to code this on R, because it is one of the main software used in the field of statistics.

3.3 Hypothesis Testing

However, once I get the maximised values for the coefficients, I need to test the significance of those coefficients. This is because I need to know if I can actually have trust in my results, and its impacts. To display how I can do this, I can give an example. In the example, let's assume that $\beta_1 = 0$ indicates that there is no association between time and the response variable Y. Therefore, I can set the null hypothesis $H_0: \beta_1 = 0$, and test this hypothesis against the alternative, that $H_0: \beta_1 \neq 0$. In statistics, we do this by using the z score to test this hypothesis. The test statistic is

$$z = \frac{\widehat{\beta}_1}{\mathrm{SE}(\widehat{\beta}_1)}, \quad (15)$$

where $\hat{\beta}_1$is the maximum likelihood estimate of β_1, and SE($\hat{\beta}_1$) is the standard error of $\hat{\beta}_1$. Two notes: (1) the hat notation in statistics refers to an estimate, and (2) the standard error of $\hat{\beta}_1$ is the standard deviation of the distribution of what $\hat{\beta}_1$ can be (Hosmer & Lemeshow, 2000). The reason why I can use the z score for hypothesis testing relates to p-values. Testing significance with p-values by setting the desired level of significance to for example, 0.05, simply tests the probability that the z score is below that level (Hosmer & Lemeshow, 2000). If the null hypothesis is rejected, then I can say that there is significant association between time and the probability of the occurrence of an inferior condition (that is, $P(Y = 1)$). Similarly, I can test the significance of the regional effects on Y. I used the R software to perform these tests.

3.4 Interpretation of the Regression Coefficients

Now, we need to figure out how to actually interpret the regression coefficients, once I get them. The best way to show this is if I illustrate it by considering the regional data model (6). For model (6) the logit function (refer to equation (9) if a refresher is needed for the logit) is

$$g(x_1, x_2, x_3) = \ln\left[\frac{P(x_1,x_2,x_3)}{1-P(x_1,x_2,x_3)}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3. \qquad (16)$$

Note that for model (6) the three predictor variables (x) are input. Now, let's say that I wish to interpret the regression coefficient β_2. If I want to only interpret that at this time, I need to fix the values of x_1 and x_3 to some value. For now, let's call $x_1 = a$ and $x_3 = b$. Now, recalling the dummy variables, we need to consider the two possibilities for x_2, that it may equal 1 or 0.

$$g(a, x_2 = 1, b) - g(a, x_2 = 0, b) = \beta_0 + \beta_1 a + \beta_2 + \beta_3 b - \beta_0 - \beta_1 a - \beta_3 b = \beta_2. \qquad (17)$$

Thus, the coefficient β_2 is the change in the logit corresponding to a one unit change in x_2, adjusting for x_1 and x_3 (Hosmer & Lemeshow, 2000). Why is it one unit of change? Because

there is one unit of difference between 1 and 0. From (16) and (17), I can now get rewrite each $g(x)$ function in natural logarithm form.

$$\beta_2 = g(a, x_2 = 1, b) - g(a, x_2 = 0, b) = \ln\left[\frac{P(a,x_2=1,b)}{1-P(a,x_2=1,b)}\right] - \ln\left[\frac{P(a,x_2=0,b)}{1-P(a,x_2=0,b)}\right]$$

$$= \ln\left\{\frac{\frac{P(a,x_2=1,b)}{1-P(a,x_2=1,b)}}{\frac{P(a,x_2=0,b)}{1-P(a,x_2=0,b)}}\right\}. \tag{18}$$

When I exponentiate with e (18) I get

$$\frac{\frac{P(a,x_2=1,b)}{1-P(a,x_2=1,b)}}{\frac{P(a,x_2=0,b)}{1-P(a,x_2=0,b)}} = e^{\beta_2}. \tag{19}$$

In (19), the first thing I can see is that the numerator $\frac{P(a,x_2=1,b)}{1-P(a,x_2=1,b)}$ is the odds for $x_2 = 1$ adjusted for x_1 and x_3. As well, the denominator $\frac{P(a,x_2=0,b)}{1-P(a,x_2=0,b)}$ is the odds for $x_2 = 0$ adjusted for x_1 and x_3. Thus, the expression e^{β_2} is actually the adjusted odds ratio (OR) comparing the Central Canada ($x_2 = 1$) with the Atlantic Canada ($x_2 = 0$) with respect to the occurrence of an inferior condition ($Y = 1$). Now, for how to analyse this, if the maximum likelihood estimate of β_2 is 0.5 for example, then the estimate of the OR is $e^{0.5} = 1.65$. Thus, this example implies that the risk of an inferior condition is 1.65 times as likely to occur in the Central Canada compared to Atlantic Canada adjusting for the other variables.

Similarly, I can also show that e^{β_3} is the adjusted OR comparing Western Canada with Atlantic Canada, and e^{β_1} is the adjusted OR for a one unit (month) increase in time.

3.4 Analysis and Results

The fit of the logistic repression model (6) to the regional data is summarized in Table 3 (see Appendix B for R codes). From here, I can answer the research questions described in Section 2.

Table 3: Fit of the logistic regression model (6) to the regional data.

Coefficients	Parameter Estimate	Standard error	p-value	*z* score	Adjusted OR
Intercept (β_0)	-1.10361	0.53684	0.0397	-2.057	0.3315716
Time (β_1)	0.10824	0.02381	5.47×10^{-6}	4.546	1.114315
Central Canada (β_2)	-1.30953	0.56229	0.0199	-2.329	0.2699469
Western Canada (β_3)	-0.45149	0.55146	0.4130	-0.819	0.6366788

Based on these results, I can now answer question a: Is the risk of an inferior condition increasing over time? For a statistic to be significant, it should have a p-value below 0.05. As the p-value for $\hat{\beta}_1$ is below 0.05, we reject the null hypothesis, meaning that time ($\hat{\beta}_1$) is significant. From here, the adjusted OR for time is 1.114315. Hence, the risk of an inferior condition increases 1.1 times for one month increase during January 2018 to January 2021. This suggests that the pandemic has significant effects to increase the risk of on inferior employment condition.

Now, I can answer question b: Does the risk of an inferior condition significantly differ by region? As $\hat{\beta}_2$ is significant, I can say that the risk of an inferior condition does significantly differ between Atlantic Canada and Central Canada. However, $\hat{\beta}_3$ is not significant. Thus, I fail to reject the null hypothesis with respect to Western Canada. As a result, I cannot determine if the risk of an inferior condition significantly differ between Atlantic Canada and Western Canada.

Question c can now be considered: Is the risk of an inferior condition higher/lower for one region compared to another region? I can only answer question c with respect to the relationship between Central Canada and Atlantic Canada, by looking at the adjusted OR. This is because Western Canada is not significant. The risk of an inferior condition due to the pandemic is 0.27 times as likely for Central Canada than Atlantic Canada adjusting for other variables. However, I can find a more intuitive conclusion by inversing the OR.

$$\frac{1}{e^{\widehat{\beta}_2}} \approx \frac{1}{0.2699469} \approx 3.704432. \quad (20)$$

Hence, I can say that an inferior condition due to the pandemic is 3.7 times more likely in Atlantic Canada than in Central Canada adjusting for other variables. This makes sense. The Atlantic provinces have a reputation in Canada of being having high unemployment rates, and being economically behind than most places in Canada, particularly Central Canada (Polese, 2020). The economic impact of the pandemic would have likely made it harder for Atlantic businesses to keep employees, than it would for businesses in Central Canada, hence the higher probability of an abnormal employment situation. On the other hand, Central Canada tends to among the more economically stable regions in Canada (Polese, 2020).

Table 4: Fit of the logistic regression model (10) to the national data.

Coefficients	Parameter Estimate	Standard Error	p-value	*z* score	Adjusted OR
Intercept (β_0)	-2.24712	0.87728	0.01042	-2.561	0.1057032
Time (β_1)	0.11443	0.04091	0.00515	2.797	1.121234

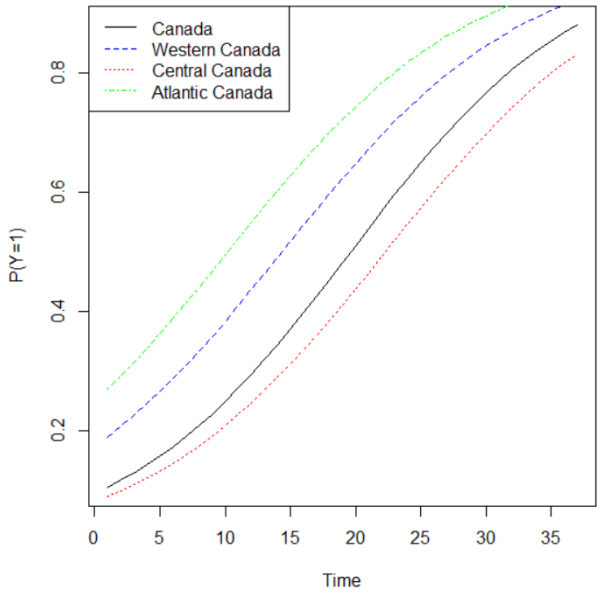

Figure 1: The risk of inferior employment condition (fitted probability values) for the national and regional data.

The last question is d: What is the overall effect of time on YOY change? To address this question, I fit the model (10) for the national data; the results are summarized in Table 4. We can see that the p-value for $\hat{\beta}_1$ is below 0.05, indicating that time is significantly associated with the risk of an inferior employment condition. The estimate of the adjusted OR is 1.1, suggesting that at the national level, the risk of an inferior condition increases 1.1 times for one month increase during January 2018 to January 2021.

Figure 1 displays the risk of inferior employment condition (fitted probability values) as a function of time for the national and regional data. The unique insight that the graphical representation gives is a visualisation of how the risk of an inferior condition increases over time. Looking at the national curve the point in time that the national curve reaches that 50% threshold is at about 24 months (in other words, January 2020). This is interesting, because the first coronavirus case in Canada was in January 2020 ("Coronavirus", 2020), with February being the month that the pandemic started to have a notable social impact in Canada ("Novel coronavirus", 2020). This solidifies the fact that there is a relationship between an inferior employment condition, and the pandemic. The figure also indicates that outsized influence that Central Canada has on the national curve. I can say this because the national curve is closest to Central Canada. This is supported by the fact that the Central Canadians make up the majority of Canada's employment workforce, resulting in an outsized influence ("Labour force", 2021).

One more thing that the graphical representation can do is provide context for the results in Table 3. One thing that was shown was that Western Canada was not statistically significant. On the graph, the curve for Western Canada is very close to Atlantic Canada. This gives a graphical justification for why Western Canada was not statistically significant. Despite the lack of significance, what the graph does show is that the increasing risk of an inferior condition in

Western Canada is more similar to Atlantic Canada's situation, than that of Central Canada. As the pandemic would hurt areas of the country with weaker economic outcomes, it makes sense that Western Canada would be further away from Central Canada, which is by far the most economically stable place in Canada (Polese, 2020). In comparison to Western Canada, Central Canada is much further away, and below the Atlantic Canada curve. This supports both the fact that Central Canada is statistically significant, but also that Atlantic Canada is further at risk of facing an inferior condition due to the pandemic.

4. Conclusion

The purpose of this exploration was to explore the impact of the pandemic on employment in Canada using a logistic regression. After analysing the results, I found that the pandemic resulted in a 1.1% increase in the change of an inferior employment condition every month. More interestingly there is indeed regional diversity in Canada, with the pandemic impacting Atlantic Canada the most, and Central Canada the least. Just on a personal level, beyond the actual analysis, this exploration expanded my horizons of what statistics is. Through this project, I have learned a great deal more about the nature of probability and just how crucial it is to understanding statistics. More than that, this paper has also expanded my understanding of why technology is needed in math. For example, I would not be able to estimate my parameters without the help of R. Simply said, this exploration has not only taught me more about statistics and math, but also about how they can reflect truths and understandings we have about the world.

Regardless, there were two principal limitations to my data set. Firstly, my data set did not include territories, which excludes an important part of Canada from my analysis. Secondly, the Statistics Canada data set rounds employment figures, so my findings are not fully exact. However, given

that my purpose was to see the impact of the pandemic in different areas of Canada, these limitations are do not invalidate my findings.

There are two extensions that I would like to conduct. First, with more time, I would like to see if there are any better data sets that do rectify the limitations in my exploration. Secondly, with a more robust knowledge of statistics, I would like to explore the Bayesian form of logistic regression. There is a long running debate in the field of statistics about whether classical statistics (which I used) or Bayesian statistics should be preferred. Seeing how Bayesian methods changes my understanding of logistic regressions, and statistics overall would be enriching.

5. References

Balaban, J. (2018, October 29). A Gentle Introduction to Maximum Likelihood Estimation. Retrieved March 27, 2021, from https://towardsdatascience.com/a-gentle-introduction-to-maximum-likelihood-estimation-9fbff27ea12f

Cieslak, J. (2021, March 08). How to Calculate Year-Over-Year Growth l Sisense. Retrieved March 27, 2021, from https://www.sisense.com/blog/calculate-year-year-growth/

Coronavirus: Here's a timeline of COVID-19 cases in Canada. (2020, March 07). Retrieved March 28, 2021, from https://globalnews.ca/news/6627505/coronavirus-covid-canada-timeline/

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression*. (2nd ed.) John Wiley & Sons Inc.

Kleinbaum, D., & Klein, M. (2010). Logistic Regression (3rd ed.). Springer.

Labour force characteristics by province, monthly, seasonally adjusted. (2021, March 28). Retrieved March 28, 2021, from https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1410028703&pickMembers[0]=3.1&pickMembers[1]=4.1&cubeTimeFrame.startMonth=12&cubeTimeFrame.startYear=2018&referencePeriods=20181201,20181201

Majaski, C. (2021, March 16). Definition of Year-Over-Year (YOY). Retrieved March 27, 2021, from https://www.investopedia.com/terms/y/year-over-year.asp

NEW POLL: COVID-19 Pandemic Having Severe Impacts on Unemployed Canadians. (2020, December 09). Retrieved March 28, 2021, from https://www.globenewswire.com/news-release/2020/12/09/2142314/0/en/NEW-POLL-COVID-19-Pandemic-Having-Severe-Impacts-on-Unemployed-Canadians.html

Novel coronavirus in Canada: Here's a timeline of COVID-19 cases across the country. (2020, March 04). Retrieved March 27, 2021, from https://www.ctvnews.ca/canada/novel-coronavirus-in-canada-here-s-a-timeline-of-covid-19-cases-across-the-country-1.4829917

Polese, M. (2020, November 13). Regional Economics in Canada. Retrieved March 27, 2021, from https://www.thecanadianencyclopedia.ca/en/article/regional-economics

What is an Odds Ratio and How do I Interpret It? - Critical Appraisal. (2020, September 18). Retrieved March 27, 2021, from https://psychscenehub.com/psychpedia/odds-ratio-2/

Appendix A

I have separated the data into national data and regional data. Appendix A consists of the national data. I will provide two things here: (1) the R code I used, and (2) the data.

I used the R code below, using the glm function, for maximising the parameters, as well hypothesis testing.

```
fit1<-glm(y~time,data=dat7,binomial(link = "logit"))
```

"dat7" refers to the data set for national data, which is shown below.

Employed Persons in Reference Month (× 1000)	Employed Persons in Current Month (× 1000)	Time (Month)	YOY Change (%)	Y
18133.3	18435.9	1	1.668753	0
18147	18459	2	1.719292	0
18163.4	18515.4	3	1.937963	0
18182.3	18501.4	4	1.755003	0
18235.2	18492.4	5	1.410459	1
18269.1	18539.4	6	1.479547	1
18290.6	18605.7	7	1.722743	0
18335.7	18552.2	8	1.180757	1
18332.8	18619.2	9	1.562227	0
18371.2	18630.9	10	1.413626	1
18439.3	18741.1	11	1.636722	0
18516.1	18741	12	1.214619	1
18435.9	18815.6	13	2.059569	0
18459	18867.3	14	2.211929	0
18515.4	18860.8	15	1.865474	0
18501.4	18963.3	16	2.496568	0
18492.4	18985.6	17	2.667042	0
18539.4	18990	18	2.430499	0
18605.7	18968.7	19	1.951015	0
18552.2	19043.9	20	2.65036	0
18619.2	19081.4	21	2.482384	0
18630.9	19078.8	22	2.404071	0
18741.1	19020.3	23	1.489774	1
18741	19074.3	24	1.778454	0
18815.6	19106.1	25	1.543932	0
18867.3	19130.3	26	1.393946	1
18860.8	18133.8	27	-3.85456	1
18963.3	16141.6	28	-14.8798	1

18985.6	16444	29	-13.387	1
18990	17385.7	30	-8.44813	1
18968.7	17802.6	31	-6.1475	1
19043.9	18016.3	32	-5.39595	1
19081.4	18388.5	33	-3.63128	1
19078.8	18482.9	34	-3.12336	1
19020.3	18537.5	35	-2.53834	1
19074.3	18484.8	36	-3.09055	1
19106.1	18272	37	-4.36562	1

Appendix B

This is the appendix for regional data. For an understanding of "Region Number", please refer back to Section 2.

The R code is below:

```
fit<-glm(y~time+factor(region),data=dat5,binomial(link = "logit"))
```

"dat5" is the data set for regional data and is listed below.

Employed Persons in Reference Month (× 1000)	Employed Persons in Current Month (× 1000)	Region Number	Time (Months)	YOY Change (%)	Y
1098.6	1100.3	1	1	0.154742	1
11123.7	11309.5	2	1	1.670308	0
5911	6026.2	3	1	1.948909	0
1091.6	1107.4	1	2	1.447417	1
11118.5	11324.3	2	2	1.850969	0
5937	6027.2	3	2	1.519286	0
1096.8	1103.6	1	3	0.619985	1
11127.1	11375.9	2	3	2.235982	0
5939.4	6035.7	3	3	1.621376	0
1096.5	1108.5	1	4	1.094391	1
11125.5	11364.4	2	4	2.147319	0
5960.3	6028.5	3	4	1.144238	1
1094.5	1106.6	1	5	1.105528	1
11158.9	11366.2	2	5	1.85771	0
5981.7	6018.6	3	5	0.616881	1
1092.7	1106.6	1	6	1.272078	1
11179.2	11404	2	6	2.010877	0
5997.2	6028.6	3	6	0.523578	1
1090.4	1112.4	1	7	2.017608	0
11209.1	11449.7	2	7	2.14647	0
5991	6043.6	3	7	0.877984	1
1089.7	1111.9	1	8	2.037258	0
11247.7	11364.1	2	8	1.034878	1
5998.2	6076.2	3	8	1.30039	1
1091.2	1115.5	1	9	2.226906	0
11265.5	11403.1	2	9	1.221428	1
5976.2	6100.6	3	9	2.08159	0
1097.2	1114.7	1	10	1.594969	0
11294.7	11422.8	2	10	1.13416	1

5979.4	6093.3	3	10	1.904873	0
1098.2	1117	1	11	1.711892	0
11339.4	11472.7	2	11	1.175547	1
6001.5	6151.4	3	11	2.497709	0
1108.7	1117.6	1	12	0.802742	1
11368.9	11488.6	2	12	1.052872	1
6038.5	6134.8	3	12	1.594767	0
1100.3	1126.9	1	13	2.417522	0
11309.5	11544.6	2	13	2.078783	0
6026.2	6144.3	3	13	1.959776	0
1107.4	1125.8	1	14	1.66155	0
11324.3	11589.4	2	14	2.340984	0
6027.2	6152.1	3	14	2.072272	0
1103.6	1130.2	1	15	2.410294	0
11375.9	11566.5	2	15	1.675472	0
6035.7	6164.2	3	15	2.128999	0
1108.5	1125.6	1	16	1.542625	0
11364.4	11648.4	2	16	2.499032	0
6028.5	6189.3	3	16	2.66733	0
1107.5	1131.1	1	17	2.130926	0
11366.2	11657.7	2	17	2.564621	0
6018.6	6196.8	3	17	2.960821	0
1106.6	1131	1	18	2.204952	0
11404	11662.8	2	18	2.269379	0
6028.6	6196.3	3	18	2.78174	0
1112.4	1124.5	1	19	1.087738	1
11449.7	11669.8	2	19	1.922321	0
6043.6	6174.4	3	19	2.164273	0
1111.9	1127.8	1	20	1.429985	1
11364.1	11743.5	2	20	3.338584	0
6076.2	6172.6	3	20	1.586518	0
1115.5	1129.9	1	21	1.290901	1
11403.1	11787.5	2	21	3.371013	0
6100.6	6163.9	3	21	1.037603	1
1114.7	1131.5	1	22	1.507132	0
11422.8	11764.4	2	22	2.99051	0
6093.3	6182.7	3	22	1.467185	1
1117	1130.4	1	23	1.199642	1
11472.7	11743.9	2	23	2.363872	0
6151.4	6146	3	23	-0.08778	1
1117.6	1129	1	24	1.020043	1
11488.6	11800.6	2	24	2.715736	0
6134.8	6144.7	3	24	0.161374	1
1126.9	1134	1	25	0.630047	1
11544.6	11832.6	2	25	2.494673	0

6144.3	6139.4	3	25	-0.07975	1
1125.8	1138.4	1	26	1.119204	1
11589.4	11844.4	2	26	2.200286	0
6152.1	6147.5	3	26	-0.07477	1
1130.2	1086.4	1	27	-3.87542	1
11566.5	11188.6	2	27	-3.26719	1
6164.2	5858.8	3	27	-4.95441	1
1125.6	966.7	1	28	-14.1169	1
11648.4	9937.3	2	28	-14.6896	1
6189.3	5237.4	3	28	-15.3798	1
1131.1	1001.3	1	29	-11.4756	1
11657.7	10121.5	2	29	-13.1776	1
6196.8	5321.2	3	29	-14.1299	1
1131	1056.5	1	30	-6.58709	1
11662.8	10737.6	2	30	-7.93291	1
6196.3	5591.5	3	30	-9.76066	1
1124.5	1067.8	1	31	-5.04224	1
11669.8	10984.6	2	31	-5.87157	1
6174.4	5750.1	3	31	-6.87192	1
1127.8	1080.5	1	32	-4.19401	1
11743.5	11157.4	2	32	-4.99085	1
6172.6	5778.5	3	32	-6.38467	1
1129.9	1095.8	1	33	-3.01797	1
11787.5	11392.2	2	33	-3.35355	1
6163.9	5900.6	3	33	-4.27165	1
1131.5	1109.6	1	34	-1.93548	1
11764.4	11411.4	2	34	-3.00058	1
6182.7	5962.1	3	34	-3.56802	1
1130.4	1128.9	1	35	-0.1327	1
11743.9	11454.3	2	35	-2.46596	1
6146	5954.3	3	35	-3.1191	1
1129	1117.9	1	36	-0.98317	1
11800.6	11439	2	36	-3.06425	1
6144.7	5928.1	3	36	-3.52499	1
1134	1125	1	37	-0.79365	1
11832.6	11187.6	2	37	-5.45104	1
6139.4	5959.4	3	37	-2.93188	1