Analysis of Employment Rate in Canada, 1981-2024

Master ECAP

Régressions Pénalisées et Sélection de Variables en Big Data

Author: Yava Vilar Valera

Date: November 6, 2024

Abstract

This study investigates the key determinants influencing the employment rate in Canada from 1981 to 2024 using a comprehensive dataset of 523 observations and 231 variables. To address the high dimensionality, machine learning techniques were applied, including penalized regressions and the General-to-Specific (GETS) model. Additionally, a regression tree method was used to explore non-linear relationships. Filtering methods like Sure Independence Screening (SIS) further refined variable selection. Results indicate that unemployment rate and lagged employment values are among the most predictive variables. The study underscores the robustness of penalized regression methods in selecting meaningful predictors amidst multicollinearity, providing valuable insights for policymakers in assessing labor market dynamics.

Contents

Introduction

Employment rate, commonly defined as the proportion of the employed population in relation to the total working age population (generally from 15 to 64 aged years)(OECD, 2024)[5](Eurostat, 2024)[1], represents a key economic indicator that reflects not only the health of a particular economy, but also the degree of performance of its labor market. While some countries experience some challenges that hinder the ability to secure optimal jobs, others possess solid regulations and a strong labor flexibility, productivity and mobility, which enhance employment opportunities and ultimately contribute to economic growth and prosperity.

Canada has historically been one of these advanced countries, characterized by a diversified array of productive sectors, ranging from natural resources extraction to technology and financial services. This diversity has enabled the country to adapt more efficiently to global market shifts compared to regions dependent on a single sector. Similarly, Canada benefits from a well-developed welfare state, with laws that protect the workers' rights, including minimum wages regulations, social security services and unemployment insurance, bolstering a stable and secure labor market. The high level of education in the country constitutes the availability and quality of human capital, which, coupled with strong demand for labor in a wide range of services such as healthcare, technology and renewed energy, helps maintain a labor demand-supply balance(OECD,2023)[4].

In the first quarter of 2024, the country added more than 78.500 jobs, entailing a higher growth rate compared to the fourth quarter of 2023, when only 55.100 new positions were established. However, Canada also faces challenges within its labor market. At the same time that new jobs are created, the unemployment rate rose from 5% in the last quarter of 2023 to 5.9% in the first quarter of 2024. Some analysts attribute this rise to differences in the growth rates of employment and population, with the latter growing at a faster pace(Actalent,2024)[3].

The study aims to pinpoint the main determinants that shape the Canadian employment rate. For this, we are using data provided by Olivier Fortin-

Gagnon *et al.* (2020) [2] which comprises 523 observations covering the period from January 1981 to July 2024, and includes 410 variables organized into eight different categories: "production, labour, housing, manufacturers' inventories and orders, money and credit, international trade and financial flows, prices and stock markets" [2].Given the high dimensionality of the dataset, the objective of this work is to identify a reduced number of key predictors using machine learning techniques, primarily penalized regressions but also GETS dimension reduction and non-linear approaches.

Chapter 1

Descriptive and Exploratoy Analysis

Missing values

Table 1.1: Missing Values in the Database

Variable	**Description**	**Missing values**
Cred_T_discontinued	Total credit	46
Cred_Household	Household credit	46
Cred_Mort	Mortage credit	46
Cred_Cons	Consumption credit	46
Cred_Bus	Business credit	46

Before conducting any econometric analysis, it is essential to undertake a descriptive and exploratory analysis to obtain a better understanding of the data. First of all, it should be specified that, among the initial 410 variables, many of them contained information related to provinces and not to the country as a whole. Having kept only those referring to Canada as a whole, the dataset has been narrowed down to 113 determinants. Furthermore, in the initial database, five variables each had forty-six missing values. Given that this implied a total of 230 missing observations, a significant number to impute, and considering the database's substantial size, the variables in

question had also been deleted. Table 1.1 identifies these regressors and summarizes the pertinent information.

Outliers

Table 1.2: Outliers in the Canadian Employment Growth Rate

Date	**Type**	**Coefficient**	**T-Stat**
March 2020	TC	-0.0618	-31.620
April 2020	AO	-0.073	-33.295
May 2020	TC	0.053	26.269
June 2020	AO	0.043	21.303
September 2020	AO	0.017	9.262
January 2021	AO	-0.012	-5.780
February 2021	TC	0.0143	7.623
April 2021	TC	-0.016	-8.896
June 2021	TC	0.016	8.789
January 2022	TC	-0.011	-5.601
February 2022	AO	0.020	9.884

After having analysed potential outliers, six TC(Transitory Change) and five AO(Additive Outliers) have been detected. Figure 5.1 in the annex plots these findings through red points in each of the outliers. Likewise, table 1.2 provides information on their emergence date, type, coefficient and t-statistic. The latter, being larger than 1.96, indicates the atypical nature of the values. The covid-19 pandemic, which originated in the beginning of January 2020 and prolonged for a few years, can explain the deviation of employment rate from usual values. Indeed, in March 2020 the Canadian labor market lost around two million jobs and the unemployment rate growth of 2.2 points was the highest recorded in the last four decades. Nevertheless, the recovery was relatively quick. By the summer of the same year, the country began to regain the jobs it had shed in a significant way and gradually maintained this pattern until September 2021, when it returned to pre-pandemic employment

levels(R.G. Jones *et al.* 2022)[6]. This can explain why there are both positive and negative coefficients throughout this period.

In order to avoid potential biases driven by the outliers' destabilizing effects, a corrected series has been made up that eliminates this inconvenience.

Stationarity

The variable of interest in the database is total employment in Canada. However, the authors Fortin-Gagnong *et al.* (2020) [2] have applied transformations to this series for it to satisfy the stationarity hypothesis. Specifically, they have differenciated the logarithm's value of the raw series. This leads to interpret the variable as the Canadian employment growth rate. The ADF(Augmented Dickey-Fuller), KPSS(Kwiatkowski-Phillips-Schmidt-Shin) and PP(Phillips-Perron) unit root tests in table 1.3 verify that the condition of stationarity is fulfilled.

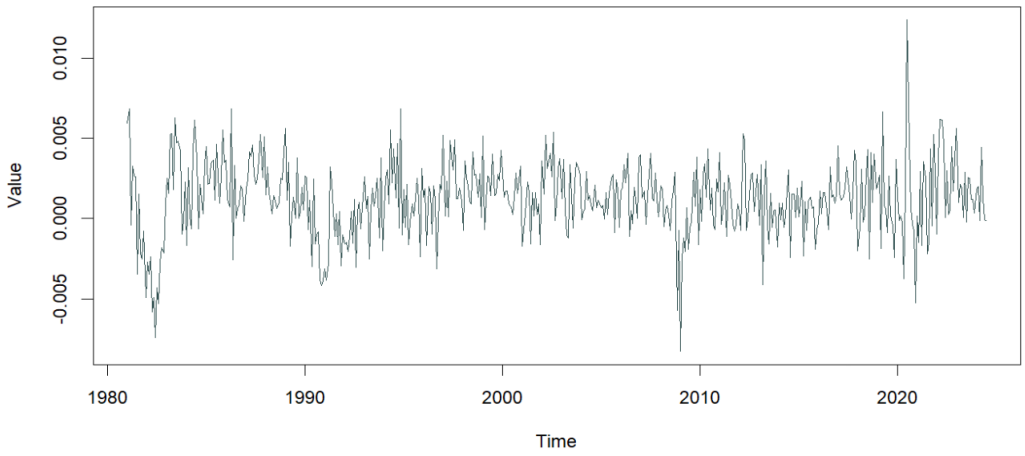
Table 1.3: Stationarity Tests

ADF		**KPSS**		**PP**	
test statistic	p.value	test statistic	p.value	test statistic	p.value
-5.76	0.01	0.083	0.1	-491.62	0.01

ADF and PP's p.value being lower than 0.05, the null hypothesis is rejected and the alternative hypothesis whereby the series is stationary is accepted at the 5% level. The KPSS test indicates the same result yet with the inverse null hypothesis. Its p.value is higher than 5% and consequently we accept the null hypothesis of stationarity.

There exist further methods to assess whether or not a series meets this requirement. Among them, Autoregressive model, order 1 = AR(1), indicates stationarity if the coefficient associated with this parameter is statistically significant and lower than 1. In the model, we obtained a coefficient AR(1) = 0.28 and a standard error = 0.042, implying significance since t = 6.7 (0.28/0.042). Thus, the series can be deemed stationary. Model results are available in figure 5.2 in the appendix.

Figure 1.1: Evolution of Employment Growth Rate in Canada, 1981-2024



Additionally, figure 1.1 displays the evolution of the employment growth rate in Canada over the last four decades (corrected from outliers). The constant tendency to return to the same mean value over time and the similarity in the deviations' width from the mean visually confirm the attainment of stationarity.

Descriptive Statistics

Table 1.4: Descriptive Statistics - Employment Growth Rate

Mean	**Median**	**S.d.**	**Skewness**	**Kurtosis**	**Shapiro test**
0.0012	0.0012	0.0023	-0.1891	1.5801	2.2e-05

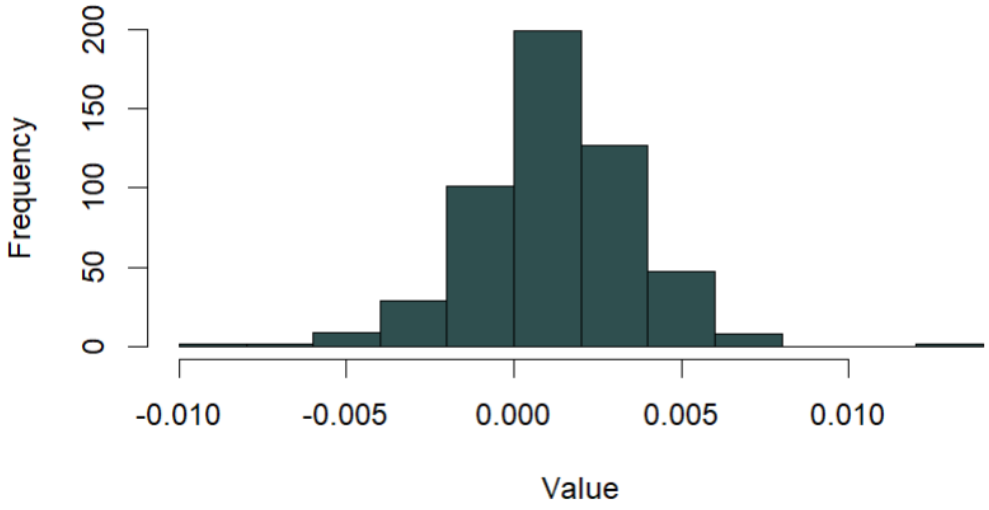
Descriptive statistics of the Canadian employment growth rate in table 1.4 indicate a positive mean of 0.0012 and a almost identical median of 0.0013. This observation drives us to deduct an accurate symmetry in the data. Yet, the median being slightly above the mean, the distribution turns out to be somewhat skewed to the left. The negativeness of the skewness' coefficient reinforces our interpretation, given that this sign exhibits the greatest presence of small values with a relatively low frequency, compared to fewer but

more frequent high values. Even so, the skewness' coefficient is rather weak, confirming the moderate asymmetry to the left.

Similarly, the wake standard deviation (s.d. = 0.0023) and the positiveness of the kurtosis' coefficient (1.5801) concordantly indicate low dispersion within the data and a concentration of the values around the mean. Indeed, a positive sign in the kurtosis test points out a peaked distribution.

On the contrary, Shapiro test's p.value being lower than 0.05, the null hypothesis of normality in the data distribution is rejected. This result indicates that even if the values' frequency in the data seem to be divided up in an almost balanced manner within the mean's left and right sides, these values are not normally distributed.

Figure 1.2: Histogram of the Canadian Employment Growth Rate



The histogram in figure 1.2 effectively reveals how there are dissimilarities in the frequency distribution between left and right, even though there seems to be a similar number of observations as a whole divided up in the two sides around the mean.

Classification

As the dataset contains a large number of variables, it is helpful to classify the data into various classes or groups that will reduce their complexity and allow an easier understanding of their relationships. For this purpose, we start by creating a PCA (Principal Component Analysis). Figures 1.3 and 1.4 show the fifteen more contributive variables to both axes 1 and 2 and axes 1 and 3. In line with this, table 1.5 provides the five more contributory variables to each of the three axes, along with their respective contribution value.

Figure 1.3: ACP - Axes 1 and 2

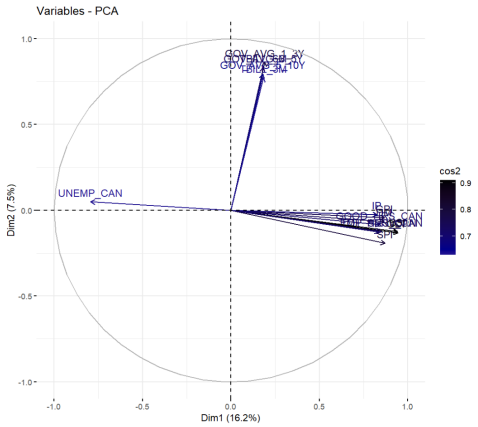
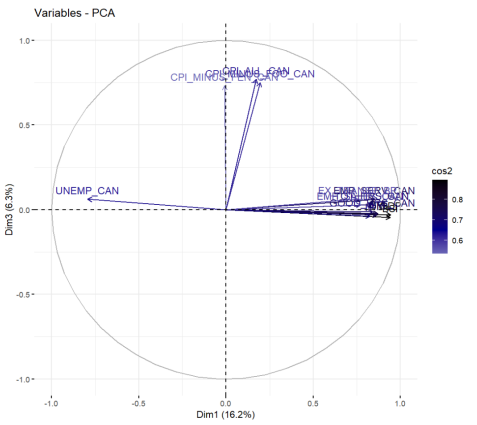


Figure 1.4: ACP - Axes 1 and 3



Based on these factors who contribute the most, we can infer the latent variables that improve interpretability. The first axis appears to relate closely to GDP or the economic wealth of the Canadian economy, as all major contributing factors are either GDP or one of its components. Additionally, these variables are plotted close to each other in figure 1.3 indicating their strong correlation and a tendency to cluster together. In contrast, axis 2 seems to be associated with government indebtedness, as it includes variables related to government bonds and treasury bills, which reflect fiscal policy and debt structure. Axis 3, on the other hand, appears linked to inflation, with contributions from indicators like the Consumer Price Index (CPI) alongside government bonds and manufacturing inventories, suggesting a connection

Table 1.5: Contribution of Variables to the Axes - PCA

Axis 1		**Axis 2**		**Axis 3**	
GDP	4.87	GOV_AVG_1_3Y	8.76	CPI_ALL_CAN	8.35
Gross Domestic Product		Governmental bonds (average rate)(1-3years)		Consumption price index	
BSI	4.84	GOV_AVG_3_5Y	8.16	CPI_MINUS_FOO_CAN	7.93
GDP Business		Governmental bonds (average rate)(3-5years)		CPI(all minus food)	
SPI	4.14	TBILL_6M	8.15	CPI_MINUS_FEN_CAN	7.55
GDP Services		Treasury bills(6 months)		CPI(all minus food and energy)	
GPI	4.11	GOV_AVG_5_10Y	7.41	G_AVG_10p,TBILL_3M	6.29
GDP Goods		Governmental bonds (average rate)(5-10years)		Government Bonds (10+years)-TBILL3M	
DM	4.09	TBILL_3M	7.01	MANU_INV_RAT	6.07
GPD Durables		Treasury bills(3months)		Manufacturing inventories to shipments ratio	

Note: Below captitalized variables rely their description. Figures represent contributions.

between inflationary pressures and production levels. Together, these factors might reveal a latent variable associated with the business cycle, because changes in government debt could signal phases of expansion or contraction, while inflation and manufacturing inventories might show shifts in demand and supply. These results indicate thus how economic wealth, government indebtedness and business cycle are related with the Canadian employment growth rate. However, it must be specified that the axes capture very little variance (16.2% the first axis, 7.5% the second axis, and 6.3% the third one), signaling weak dimension reduction effectiveness and possible noise dominance.

For further exploration, a hierarchical cluster analysis is available in figure 5.3 in the annex to illustrate how data are grouped at different levels of similarity or dissimilarity. With an optimal partition in 2 classes (k=2), the dendrogamme shows the data likely has two defined groups that are relatively distinct from each other. This is because the two main branches split at a relatively high height (more than 100), which indicates a significant level of dissimilarity between them. Below one of the two primary branches, there are numerous smaller branches, representing subclusters within the main

cluster. Given that they merge at lower heights, they are more similar to each other. Noticeably, there is a big difference in the quantity of observations belonging to one or the other of the two major clusters, potentially indicating atypical values in the one containing very few data. Besides, the quality of the partition is very low (0.10), suggesting that classes are not well-defined nor meaningful. This is why this analysis won't be continued.

Correlations

It is now essential to assess the correlation level within the data in order to identify potential problems that might arise from collinearity. For this, in a first instance we are going to evaluate the correlation between the explanatory variables and then between the dependent variable and the independent ones. For legibility and simplicity reasons, the correlation matrix from figure 1.5 has filtered all regressors who are correlated at least at 0.9 level with another regressor. This does not mean, though, that within the whole matrix there are not correlations of lower magnitude. It can be seen how there are six variables presenting correlations higher than 0.9, which depicts multicollinearity. To mention some examples, GDP (Gross Domestic Product Total) and GPI (GDP Goods) present a correlation of 0.98, and GPI and IP (GDP Industrial Production), correlate at 0.93. Besides, even more variables are correlated at 0.8 or more. Thus, these results underline the importance of considering appropriate approaches and conducting a careful analysis.

Figure 1.5: Correlation Between Explanatory Variables

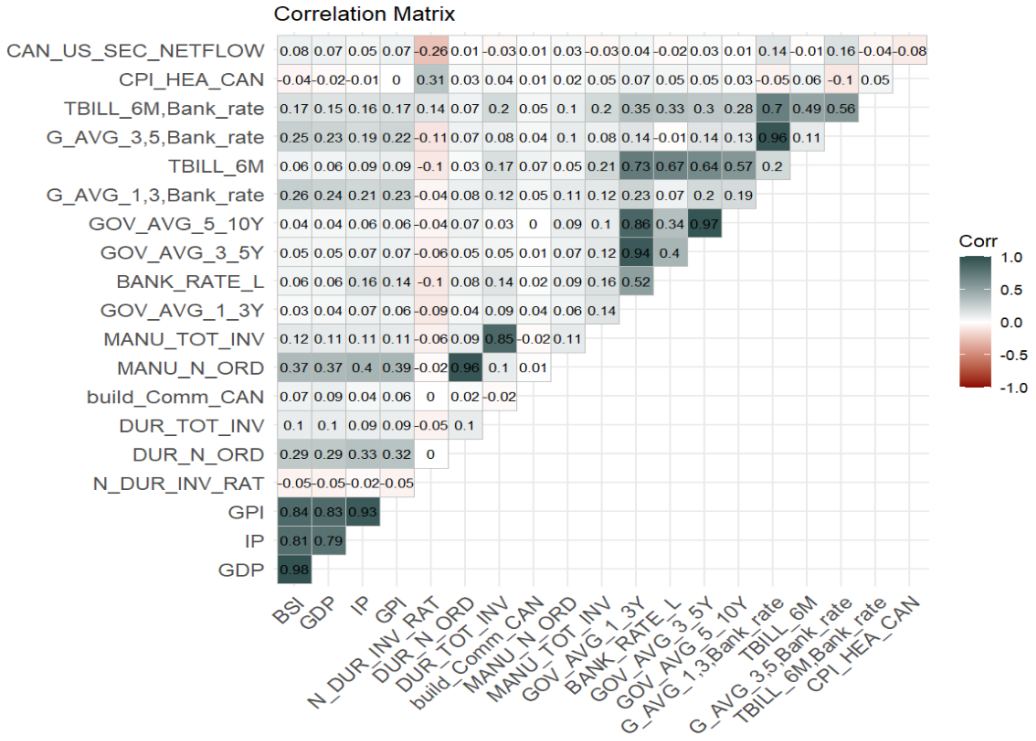


Figure 1.6: Correlation of Y with X Variables

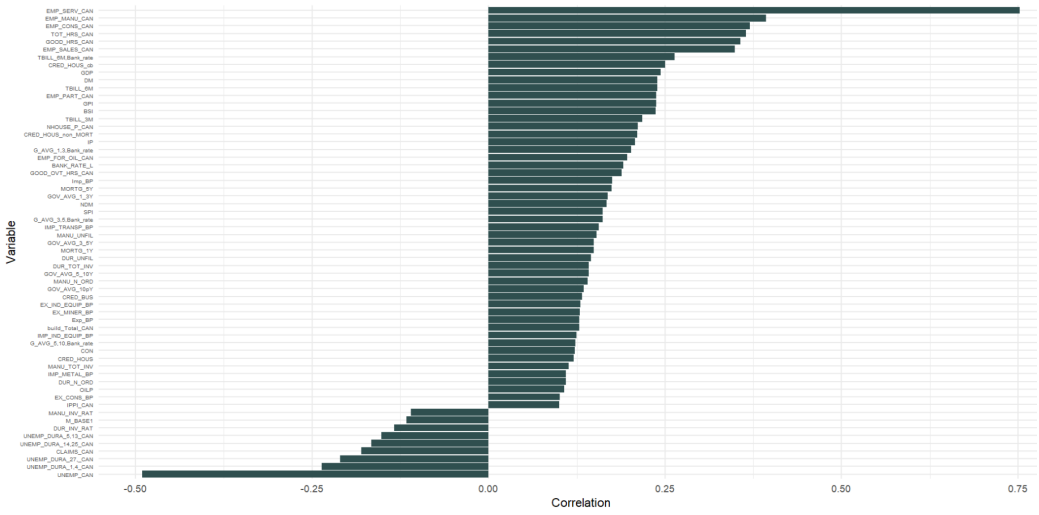


Figure 1.6 reflects correlation levels between the employment growth rate and each of its determinants that exceed 0.1 or are below -0.1. This threshold was chosen to enhance readability without excluding a large number of factors by setting a higher cutoff. Most predictors exhibit from low to moderate correlation with the dependent variable. In fact, only Emp_Serv_Can (Employment services) surpasses a correlation of 0.7, and the remaining variables don't exceed 0.5 in neither the negative nor the positive direction. Unemp_Can (Unemployment rate) is the only variable presenting a relatively high negative correlation of -0.5, which makes sense because the more unemployment grows, the more we can expect that employment will decrease and vice versa. The factors showing the strongest correlations with the dependent variable are likely to be the most significant candidates for selection within a large dataset.

Chapter 2

Variable Selection

Given the large dimensionality of our database, and the resulting inconveniences bound to arise with traditional econometric tools such as the Ordinary-Least-Squares linear regressions, it is advantageous to make use of machine learning techniques, such as penalized regressions, aimed at avoiding over-fitting and selecting only the most important and useful predictors explaining the Canadian employment rate. This will considerably reduce the model's complexity and with it the variance associated. The approach we will be covering in this work includes model's selection technique through GETS method, and five penalized regressions: LASSO, Elastic-Net, SCAD, Adaptative-LASSO, and Ridge. While the four first regressions will select a few relevant contributors, the last one will only reduce the coefficients' value of the least pertinent variables without leading to a variables' selection.

Before diving into this, it must be said that the dependent variable has been lagged for four periods and the regressors for one period, which is expected to capture trend dependencies that occur in different time scales and decrease the noise in the data. This procedure has led to end up with a database composed of 231 variables, the original variables added to the lagged ones.

GETS

GETS (General-To-Specific) method of variable selection is a powerful technique for obtaining simplified and interpretable models. Through a top-down approach, it eliminates successively non-significant predictors whose contribution is very limited, resulting in a parsimonious model. When applying this method to our data, we could encounter an issue related to the potential singularity rather than invertibility of the regressors matrix, which could arise from collinearity; and with the high number of explanatory variables. However, the GETS modeling was successful and 100 variables were assessed as relevant, as illustrated in table 2.1. In a subsequent section, we will reduce the big dimensionality of our dataset by implementing the Sure Independence Screening (SIS) filtering, and retry the GETS modeling with an attempt to narrow down the final variables chosen by the model.

Table 2.1: Results of the GETS Method

Method	Variables
GETS	100

LASSO

The penalized regression LASSO, as displayed in table 2.2, has selected two variables with a λ equal to 0.32.

Table 2.2: Results of the LASSO Method

λ	Variables
0.32	2

At first glance, we can say that the λ value could be high enough to justify the exclusion of so many variables. Nevertheless, the multicollinearity among the data may drastically reduce the number of variables retained without the need to have a particularly high λ. Indeed, in the presence of near-perfect collinearity, the LASSO model tends to select only one regressor

from a group of correlated variables. However, these other factors might still play a significant role in explaining the phenomenon of interest. This is one of the reasons for exploring alternative regression methods with different penalty structures, which aim to address this limitation.

Elastic-Net

The Elastic-Net method acts as a bridge between Ridge and LASSO. While Ridge regression uses the Euclidean norm (L2) with an α of 0, LASSO employs the L1 norm with an α of 1. Elastic-Net combines both L1 and L2 norms, and the value of α can range between 0 and 1. In this project, three approaches were applied to determine the Elastic-Net α. In the first approach, a fixed α of 0.5 was manually assigned. For the other two methods, an optimal value of α was selected using grid and random searchs.

Table 2.3: Results of the Elastic-Net Method

Pre-fixed value		Grid search		Random search	
$\alpha = 0.5$		$\alpha = 0.83$		$\alpha = 0.2$	
λ	Variables	λ	Variables	λ	Variables
0.46	6	0.32	3	0.32	42

Table 2.3 shows the outputs of the three approaches, which appear somewhat inconsistent when comparing the strength of the penalty (captured by λ), its connection with α, and the number of selected variables. When comparing grid to random search, two quite different α have led to the same λ of 0.32, which can seem counterintuitive, as we would expect a smaller α (favoring L2 regularization) to be associated with lower λ values and, consequently, less stringent penalties. What is more, the number of variables differs considerably from one method to the other (from 3 to 42). We can guess random search has favored a more Ridge-like penalty and grid search a LASSO-like penalty, aligned with their respective α and the final number of regressors selected, in spite a same λ. Finally, an α of 0.5 amidst the other two values has yield the highest λ of 0.46, a strong penalty which can be deemed consistent with the high number of coefficients being shrunk to zero.

SCAD

Table 2.4 provides results obtained from SCAD regression. We can see that fewer variables have been excluded compared to the LASSO method (10 compared to 2), which might stem from the lower degree of penalty applied (0.13).

Table 2.4: Results of the SCAD Method

λ	Variables
0.13	10

Additionally, SCAD modeling is advantageous when collinearity among the data prevails, as it employs a smooth penalty that penalizes the largest coefficients less. This means that it can retain some important determinants that LASSO would drop.

Adaptative-LASSO

The Adaptative-LASSO regression, which has been computed using Ridge as first fit, is aimed at correcting some limitations associated with LASSO method. We can expect more reliable results given that the Adaptative-LASSO approach fulfills the oracle property, whereby the model distinguishes non-relevant and relevant variables and consistently assigns zero or non-zero coefficients. By using a weighted penalty in contrast to a uniform one, the most pertinent predictors will receive the weakest penalties, while the inverse applies to the less relevant factors.

Table 2.5: Results of the Adaptative-LASSO Method

λ	Variables
0.91	0

In table 2.5, λ coefficient is 0,91, much higher than the obtained in previous approaches, perhaps explaining why all variables have been excluded by the model.

Ridge

Finally, Ridge regression adds a penalty to the size of the coefficients in a uniform manner, making it an optimal model to address collinearity problems that OLS fails to address, yet it is not designed for variable selection, which explains why all variables have been retained in table 2.6.

Table 2.6: Results of the Ridge Method

λ	Variables
2.98	230

Regarding λ coefficient, it is equal to 2.98, indicating a very strong penalty on the coefficients, and therefore, their potential individual low weight in predicting the variable of interest.

Chapter 3

Variable Selection After Filtering

We now opt for reducing the dimensionality of our large database and then to re-implementing all methods and regressions considered thus far. This procedure is appropriate in the context of our analysis because we can expect a reduction in collinearity among regressors as well as a decrease in data noise. Besides, narrowing down the high dimensionality can help avoid overfitting. As mentioned earlier, we apply SIS filtering, a screening approach adapted to linear models. 65 regressors have been selected by the model.

GETS - SIS

GETS modelling, when implementing it alongside ARCH test for robustness, led to some diagnostic errors related to model checks, which could be arise from autocorrelation of residuals, hereroskedasticity, normality of residuals or overfitting, among others . Therefore, it has been modeled in a subsequent attempt without ARCH test, leading to 22 variables, as shown in table 3.1 , a considerable reduction with regard to GETS without SIS filtering.

Table 3.1: Results of the GETS Method - SIS

Method	Variables
GETS	22

LASSO - SIS

Table 3.2 displays the results obtained from LASSO after having applied SIS filtering.

Table 3.2: Results of the LASSO Method - SIS

λ	Variables
0.32	2

We can appreciate how an identical number of variables has been retained along with the same λ set at 0.32 compared to LASSO before SIS. It remains to see, in a subsequent section, whether the selected variables are the same. If that is the case, the result may suggest that these predictors are the really paramount ones while the remaining have a weak predictive power. Indeed, if SIS has truly improved multicollinearity issues, LASSO after this filtering has not picked more variables.

Elastic-Net - SIS

Table 3.3 provides results concerning Elastic-Net penalized regression. As previously, three approaches have been considered to choose α value.

Table 3.3: Results of the Elastic-Net Method - SIS

Pre-fixed value		Grid search		Random search	
$\alpha = 0.5$		$\alpha = 0.88$		$\alpha = 0.1$	
λ	Variables	λ	Variables	λ	Variables
0.56	2	0.38	1	0.33	36

Starting with fixed 0.5 α, it has led to a high λ (0.56), retaining only 2 variables, suggesting a parsimonious approach. Grid search with α at 0.88 and λ at 0.38 retains just 1 variable, focusing on the most predictive factor. In contrast, random search with a lower α of 0.1 and λ of 0.33 leads to 36 variables, favoring a broader model that includes more features, yet it might risk adding noise. Again, grid search appears to have adopted a LASSO-like

strategy while random search a Ridge-like. We can see from these models how higher λ values have increased sparsity, selecting only the most predictive variables, whereas a lower α value has reduced the penalty coefficient, with many more regressors remaining in the model.

SCAD - SIS

Results concerning SCAD method are displayed in table 3.4.

Table 3.4: Results of the SCAD Method - SIS

λ	Variables
0.18	3

We can see how few variables have been maintained when comparing it to a situation without filtering (3 compared to 10), at the same time that λ is greater (0.18 compared to 0.13). SIS filtering reflects a more focused set of predictors that are directly relevant to the outcome being modeled, resulting in fewer variables being retained and the penalized regression working more effectively in the selection.

Adaptative-LASSO - SIS

According to table 3.5, Adaptative-LASSO under screening filtering has maintained one predictor alongside a strong value of λ set at 0.59.

Table 3.5: Results of the Adaptative LASSO Method - SIS

λ	Variables
0.59	0

In spite a lower λ value than Adaptative-LASSO without SIS (λ=0.91), the model continues shrinking all coefficients to 0.

Ridge - SIS

As expected, Ridge did not exclude any explanatory variable but kept all of them, as shown in table 3.6. Regarding λ, it has been established by cross validation at 2.98, a very strong penalization which is assumed to shrink the coefficients near zero, resulting in a more stable model that may limit the influence of predictors.

Table 3.6: Results of the Ridge Method - SIS

λ	Variables
2.98	65

Chapter 4

Non-linear approach: Regression Tree

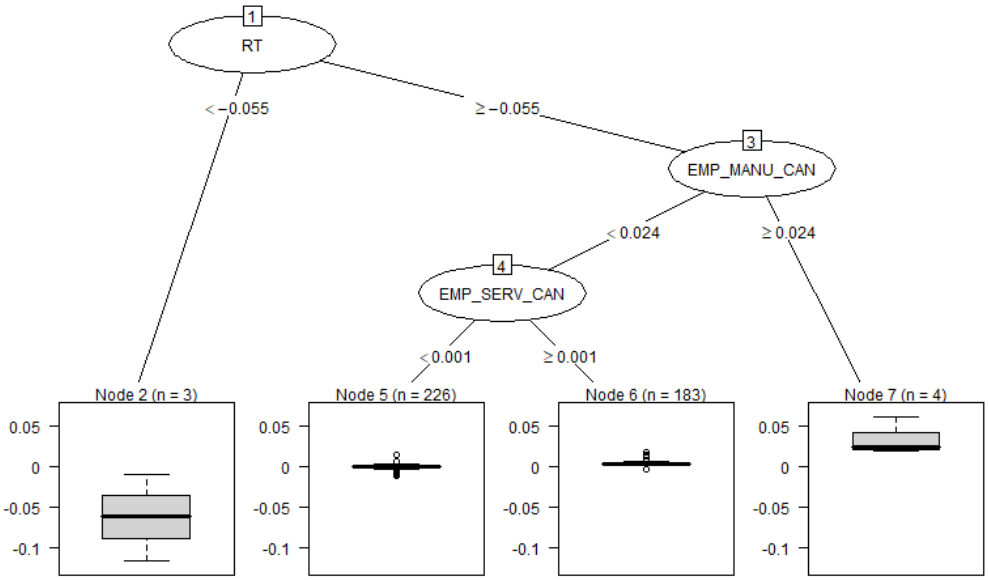
Besides all linear models that have been designed in the analysis thus far, it is worth testing non-linear techniques given that these can be appropriate in a context where high dimensionality leads to multiple kinds of complex relationships between variables beyond linear ones. Non-linear models offer flexibility and adapt effectively to different data structure scenarios. Some of these methods include ensemble methods such as Gradient Boosting, parallel methods such as Random Forest, neuronal networks, and decision trees, among others.

In this analysis, a regression tree has been modeled, a machine learning technique guided by predictability and accuracy whose main goal, in addition to its ability to cluster data, is to make predictions on continuous variables. Essentially, the tree splits into several dichotomies based on a decision rule, in such a way that by following each of the tree branches guided by the dichotomies, we obtain a final prediction on the outcome of interest. The tree is composed of various nodes, going from the root node containing the most relevant variable, to the terminal nodes that reveal prediction values.

The regression tree has been made with caret and rpart packages in R. In a first instance, data has been distributed into a train and test game aimed at evaluating the performance of the model. Then, two modeling options have

been compared by modifying minbucket (minimum number of observations that must exist at each terminal node) parameter. In the first model, a minbucket of 10 was specified. This led to choose, by cross-validation, 4 terminal nodes corresponding to a 0.0011 cp (complexity parameter). With an attempt to allow the model become more complex and capture more details, minbucket parameter was decreased to 3, leading again to 4 terminal nodes but with a slightly lower complexity captured by a of 0.0015 cp. When evaluating the RMSE (Root Mean Squared Error) on the predictions made in the data test, the second model exhibited higher prediction performance (0.001 RMSE compared to 0.004), the reason why this is the version presented in the study. Figure 4.1 displays the regression tree revealing the decision scheme along with the most important predictors.

Figure 4.1: Regression Tree on the Canadian Employment Growth Rate



First of all, the importance of the variable RT (GDP Retail Trade, expressed as the first difference of its logarithm) stands out. Should the growth rate of the Canadian retail trade be lower than -0.055, employment growth rate will be, as an approximate average, -0.06. Given that the boxplot is

rather large, we can nevertheless expect a variety in the values taken by the Canadian employment growth rate, ranging from -0.03 to -0.09. This observation constrasts to the outputs of nodes 5 and 6. If the Canadian retail trade growth rate is larger than 0.055, and the employment manufacturing growth rate (first difference of EMP_MANU_CAN's logarithm) smaller than 0.024, the employment growth rate will be very close to 0 regardless the employment services growth rate (first difference of EMP_SERV_CAN's logarithm) be above or below 0.001. Additionally, we can see that the majority of observations belong to either of these two categories. Finally, a positive average of employment growth rate is predicted in the instances where Canadian retail trade is lower than -0.055 and employment manufacturing growth rate (first difference of EMP_MANU_CAN's logarithm) higher or equal than 0.024. Table 4.1 shows additional relevant variables and its relative importance scaled to 100.

Table 4.1: Contributive Variables in the Regression Tree

Variable	**Description**	**Importance**
RT	GDP Retail Trade	17
BSI	GDP Business	15
GDP	GDP Total	15
SPI	GDP Services	15
GPI	GDP Goods	11
EMP_MANU_CAN	Employment Manufacturing	5
UNEMP_DURAvg_CAN	Unemployment Average Duration	5
EMP_SERV_CAN	Employment Services	2

The same procedure has been applied with the reduced database guided by SIS. While the regression tree is exactly the same as that without shrinkage, and presents the same error levels, NDM predictor has become more relevant with a relative importance of 4, replacing UNEMP_DURAvg_CAN. The new list of paramount regressors is shown in table 4.2. The fact that practically the same factors are retained may indicate a robustness to high dimensionality in our context. Indeed, the model has been able to recognize those factors that contribute the most with a high initial number of variables,

and reducing dimensionality has not brought about significant modifications.

Table 4.2: Contributive Variables in the Regression tree - SIS

Variable	**Description**	**Importance**
RT	GDP Retail Trade	17
BSI	GDP Business	15
GDP	GDP Total	15
SPI	GDP Services	15
GPI	GDP Goods	11
EMP_MANU_CAN	Employment Manufacturing	5
NDM	GDP Non Durable Goods	4
EMP_SERV_CAN	Employment Services	2

Chapter 5

Summary of All Models

Table 5.1 summarizes the outputs obtained with each of the models before and after filtering. Strikingly, it can be seen that all penalized regressions with dimensionality reduction have selected a lower number of variables post-filtering. This may be a common result, on the one hand, because by starting with a smaller set of predictors, penalized regressions have a smaller search space, thereby making it likely that the final models end up with fewer variables. Besides, after SIS, regularizations add a second filter, further excluding other predictors that do not provide any value when combined with other variables.

We can ask ourselves whether the mitigation of multicollinearity was effective by SIS filtering. Indeed, prior to this reduction, considering that there were variables strongly correlated with each other, or variables that disturbed and made selection difficult, the models may have been forced to discard relevant regressors that, in cleaner situations, could have been identified as such. For this reason, we would expect more variables to be retained after filtering, specially in regressions such as LASSO. Nevertheless, the opposite happened, and it can be due either because collinearity remains or because certain variables are actually still considered non-contributive. In all cases, we may prefer selection under SIS, given the intrinsic advantages this adoption implies.

Another remarkable point is that, when comparing before and after SIS

Table 5.1: Summary of the Variables Selected by the Models

Method		**Without SIS**			**With SIS**		
		α	λ	Variables	α	λ	Variables
GETS				100			22
LASSO			0.32	2		0.32	2
	Pre-fixed	0.5	0.46	6	0.5	0.56	2
Elastic-Net	Grid search	0.83	0.32	3	0.88	0.38	1
	Random search	0.2	0.32	42	0.1	0.33	36
SCAD			0.13	10		0.18	3
aLASSO			0.31	0		0.59	0
Ridge			2.98	230		2.98	65

within the same models, λ is always higher (except for LASSO) and the number of variables lower. This is coherent, since, as it has already been said, an increased λ penalizes stronger. This can be an additional reason explaining why there are fewer predictors after SIS. However, if we compare across models, the link between the value of λ and the final variables does not hold. For instance, Elastic-Net random search before SIS, with a λ of 0.32, selected 42 variables, while SCAD, with a λ of 0.13, just 10. Effectively, λ values themselves are not directly comparable across penalized methods due to each model's unique approach.

Table 5.2 presents the coefficients for the variables retained by each model analyzed in this study before filtering, and table 5.3 after filtering. Additionally, an Ordinary Least Squares (OLS) regression has been computed, whose variables, for consistency and ensuring comparability, have been standardized. Notably, due to the large number of variables selected by Elastic-Net with random search and GETS, as well as the natural inclusion of all regressors by OLS and Ridge, only the five most contributive coefficients are

Table 5.2: Comparison of Coefficients Across Models

Variable	**OLS**	**Ridge**	**EN** α=0.5	**EN** G.S	**EN** R.S	**SCAD**	**LASSO**	**aLASSO**	**GETS**
GDP	**2.601***								**2.103**
GPI									-1.214
SPI	**-1.913****								**-1.586**
EMP_SERV_CAN									1.007
EMP_FOR_OIL_CAN		0.037			**0.078**	0.057			
EMP_MANU_CAN					0.051				
UNEMP_CAN		-0.034	-0.101	**-0.100**	-0.060	**-0.288**	**-0.059**		
UNEMP_DURA_1.4_CAN					-0.054				
UNEMP_DURA_27._CAN		-0.038			-0.070				
M_BASE1						0.051			
EMP_CAN_lag1			0.003						
EMP_CAN_lag2		**0.043**	**0.059**	0.045		0.142	0.002		
NHOUSE_P_CAN_lag1		0.041	0.036	0.016		0.100			
N_DUR_INV_RAT	1.599*								1.023
DM_lag1			0.014						
N_DUR_INV_RAT_lag1	-1.800*								
DUR_INV_RAT_lag1	-1.848.								

Note: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; . $p < 0.1$ G.S = Grid Search; R.S = Random Search

Table 5.3: Comparison of Coefficients Accross Models - SIS

Variable	**OLS**	**Ridge**	**EN** α=0.5	**EN** G.S	**EN** R.S	**SCAD**	**LASSO**	**aLASSO**	**GETS**
BSI									-0.055
EMP_FOR_OIL		0.041			0.082				
EMP_CONS_CAN					0.068				
EMP_SERV_CAN									0.443
G_AVG_1.3.	0.52 *								0.412
UNEMP_CAN		-0.039	**-0.075**	-0.039	-0.092	**0.267**	**-0.059**		
BSI_lag1	0.98*								**1.286**
GDP_lag1	**-2.04***								-1.172
SPI_lag1	0.63*								
GPI_lag1	0.55*								
EMP_CAN_lag1		0.038			0.050	0.016			
EMP_CAN_lag2		**0.054**	0.032		**0.122**	0.173	0.002		
EMP_CAN_lag3		0.035							

Note: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; . $p < 0.1$ G.S = Grid Search; R.S = Random Search

displayed for all approaches. At the end of the document, on the contrary, a complete list of variables along with their coefficients is available for SCAD and Elastic-Net random search.

When interpreting OLS results, a substantial number of significant determinants affect Canada's employment growth rate, even though only five of them are shown in the tables. Particularly influential is GDP with a positive impact without SIS: if GDP increases by 1%, the employment growth rate is expected to increase by 2.60%, significant at the 5% level. In contrast, SPI (GDP Services) and NDUR_INV_RAT (Manufacturing inventories to shipments ratio) have a negative effect at the 1% and 5% levels respectively; specifically, a 1% increase in GDP services and manufacturing inventories to shipments ratio would respectively lead to a 1.91% and 1.6% decrease in the employment growth rate. In contrast, when applying SIS, we notice that, is some cases, the most influential variables turn out to be the same but this time lagged for one period and with an inverse sign, as it arises with GDP_lag1 and SPI_lag1.

When comparing the three approaches of Elastic Net with each other, we observe that the same determinants are often selected, with random search being the method who selects higher numbers of predictors. However, even if similar, the importance assigned to each of them is not exactly the same, particularly between pre-fixed 0.5 α and random search. In particular, the largest coefficients do not belong to the same variable. Contrastingly, there is a higher tuning between pre-fixed 0.5 α and grid search, because within the more restricted selection this latter makes, it retains, in order of importance, the variables presenting the largest coefficients among those selected by pre-fixed 0.5 α.

Regarding LASSO and Adaptative-LASSO, these have turned out to be the most choicy methods. It stands out how Adaptative-LASSO carries out a stronger variable selection. Coefficients retained by LASSO are so small that Adaptative-LASSO directly shrinks them to 0.

On the other hand, in many cases, we could expect a similarity and a correlation in the coefficients between Ridge and OLS. However, in this study this does not appear to be true, at least in what concerns the largest coefficients. Indeed, none of the five most contributing variables matches. In contrast, SCAD and Ridge appear to be more alike, as well as OLS and GETS.

Overall, we can ask ourselves whether coefficients remain unchanged before and after SIS. For LASSO, this is the case, but not for the other methods, although their differences are not very pronounced. Likewise, we observe that the same variables consistently emerge as the most relevant across models. Specially, in the most selective methods, UNEMP_CAN (Unemployment growth rate) and EMP_CAN_lAG2 (Employment growth rate lagged for two periods) always emerge.

When flexibilizing regularization and contemplating alternative methods, this allows for the selection of additional variables which also match across these other methods. This consistency supports the robustness of the models and strengthens the reliability of the results, allowing for confident identification of key determinants.

However, we observe a distinction between OLS and GETS, and the penalized regressions. The largest coefficients in the firsts are not consistently aligned with selection from penalized regressions. This can be explained by the inherent differences in their objectives and approaches. In fact, while the magnitude of OLS coefficients reflects the relationship between each predictor and the outcome, the penalized regression selects variables based on their contributions to predictive accuracy within the context of the entire model. Thus, a variable with a large OLS coefficient may not contribute significantly when considering its impact alongside other predictors. What is more, in the presence of multicollinearity, OLS can produce large coefficients for correlated predictors. Nevertheless, penalized regression techniques can mitigate the impact of multicollinearity by selecting variables and reducing their coefficients. Similarly, penalized regression seeks a balance between bias and variance, while OLS and GETS prioritize reducing error.

It is also remarkable how the more contributive factors in the regression tree differ from selection by penalized regressions. This might be due to the non-linear interaction in regression trees where certain variables can be relevant only in specific sub-regions of the data, whereas penalized regressions only capture linear relationships between the predictor and the outcome.

Finally, to pinpoint the best model, it would be necessary to forecast the outcome of interest and compare errors across models. Still, based on our

data and particular context, we can discard methods such as LASSO, given its ineffectiveness in dealing with collinearity; Ridge if our goal is to select predictors and obtaining a parsimonious model; Adaptative-LASSO because of its strong penalization and the resulting zero predictors; and GETS because none of the most powerful factors matches with selection from penalized regressions. Thus, both SCAD and Elastic-Net seem to be optimal alternatives. Elastic-Net could be even preferred because it offers a powerful combination between LASSO and Ridge, and compared to SCAD, is more effective including groups of correlated variables rather than choosing only one from each correlated group.

Conclusion

In conclusion, this paper aimed to identify the most relevant factors for explaining and predicting the Canadian employment growth rate from a diverse set of regressors by employing machine learning techniques, including dimension reduction modeling, penalized regressions, and a non-linear regression tree approach. A consistency in results within penalized regressions is noted, with the unemployment growth rate and the employment growth rate lagged for two months being the variables exercising the largest predictive influence on the outcome of interest. With few exceptions, the scarce factors retained by the models were aligned with their respective high λ values. Equally important, SIS filtering led in all cases to an increased number of predictors being shrunk to zero, indicating that the model became more parsimonious and focused on the most significant variables. Useful results provided by this analysis can be extrapolated to real-life situations in which understanding the relationship between the Canadian employment growth rate and the selected predictors is necessary. For further research, it would be valuable to explore additional predictors or alternative modeling techniques to enhance the predictive power and generalizability of the findings.

Bibliography

[1] Eurostat. Gossary: Employment rate. URL:https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Employment_rate, 2024.

[2] Olivier Fortin-Gagnon, Maxime Leroux, Dalivor Stevanovic, and Stéphane Surprenant. A large canadian database for macroeconomic analysis. *Canadian Journal of Economics/Revue canadienne d'économique*, 55(1799-1833), 2020.

[3] Eliza Hetrick. Labour market economy report: Quarter one 2024. Technical report, Actalent, 2024.

[4] OECD. canada , dans oecd employment outlook 2023 : Artificial intelligence and the labour market. DOI:https://doi.org/10.1787/f2d118bd-en, 2023.

[5] Organization for Economic Cooperation and Development. Employment rate. URL:https://www.oecd.org/en/data/indicators/employment-rate.html#:~:text=Employment%20rate%20is%20the%20extent,to%20work)%20are%20being%20used., 2024.

[6] Stephan R.G. Jones, Fabian Lange, W.Craig Riddell, and Casey Warman. The great canadian recovery: The impact of covid-19 on canada's labour market. Technical report, IZA. Institute of Labour Economics, 2022.

Annex

Figure 5.1: Outliers in the Evolution of Employment Rate in Canada

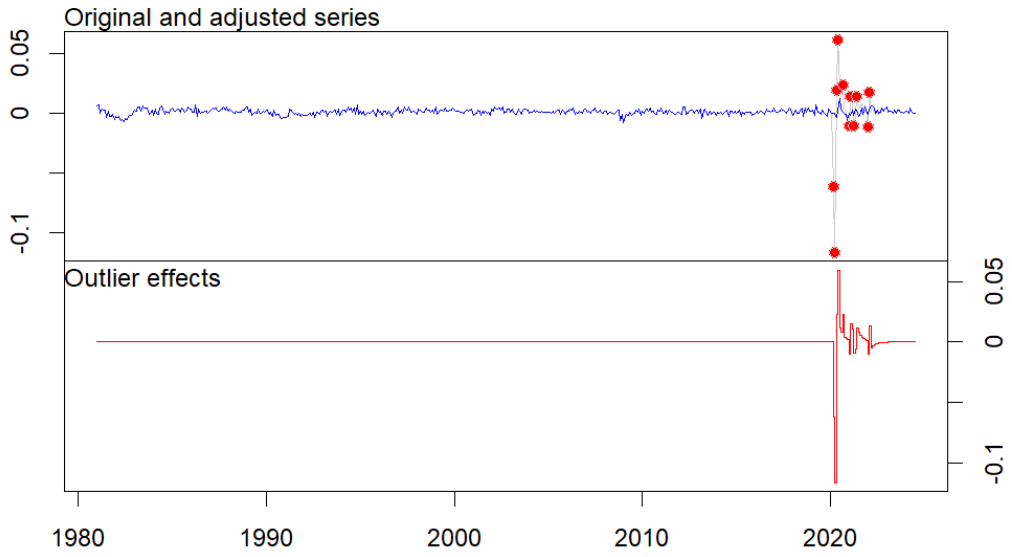


Figure 5.2: Model AR(1) - Canadian Employment Growth Rate

```
Call:
arima(x = adj, order = c(1, 0, 0))

Coefficients:
         ar1  intercept
      0.2856     0.0012
s.e.  0.0420     0.0001

sigma^2 estimated as 4.945e-06:  log likelihood = 2452.64,  aic = -4899.28

Training set error measures:
                       ME        RMSE         MAE  MPE MAPE      MASE       ACF1
Training set -3.115407e-06 0.002223716 0.001709166 -Inf  Inf 0.7995488 -0.0782339
```

Figure 5.3: Dendrogramme with 2 classes

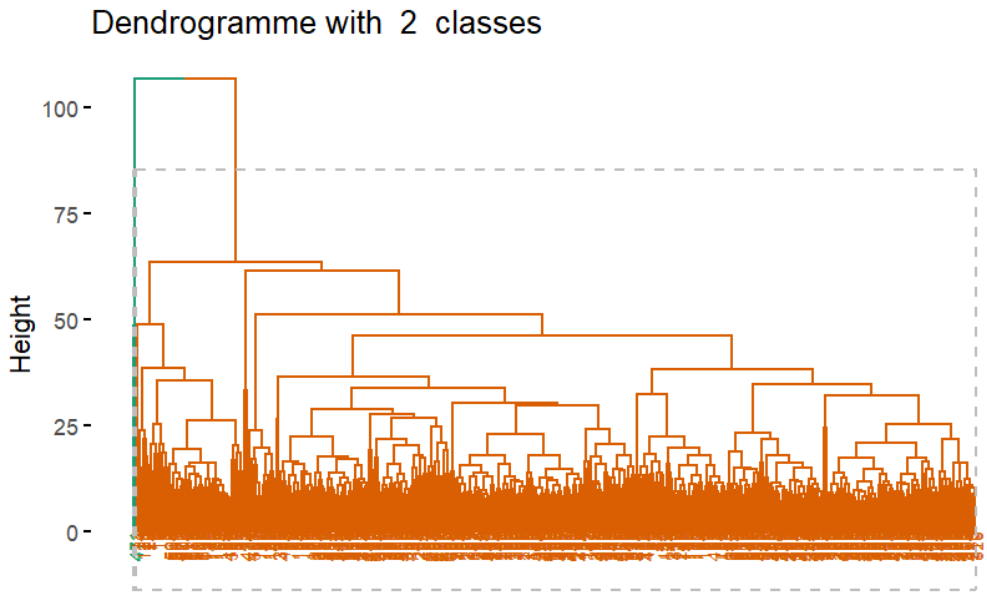


Figure 5.4: Best λ selection, LASSO

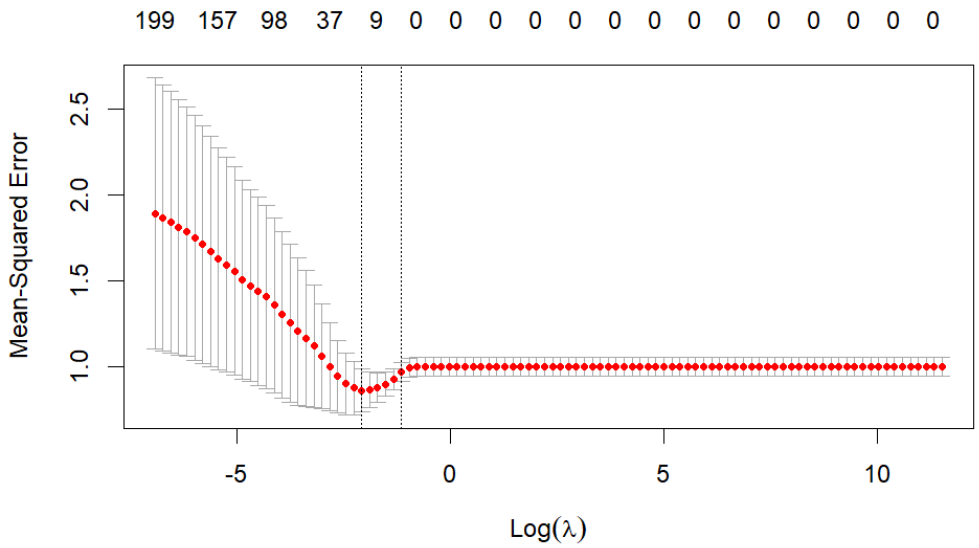


Figure 5.5: Best λ selection, Elastic Net (0.5)

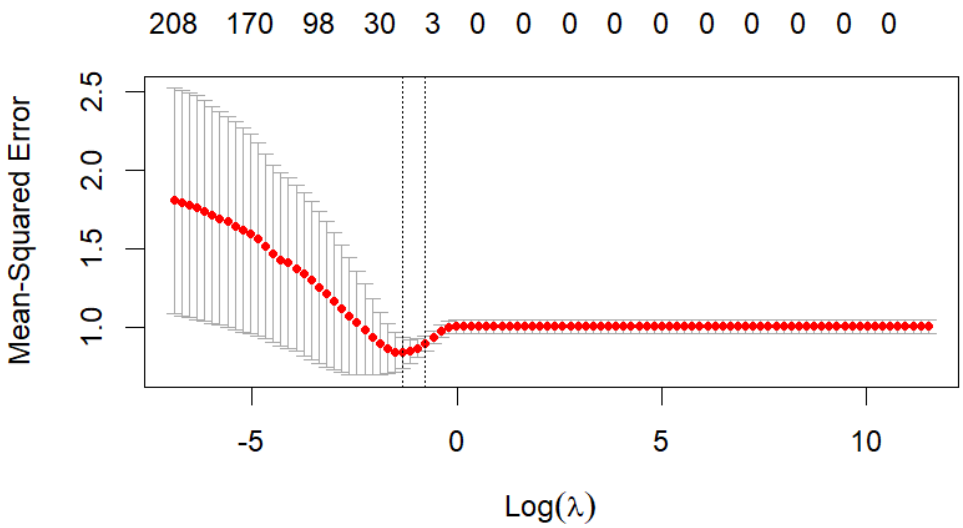


Figure 5.6: Best λ selection, Elastic Net (Grid Search)

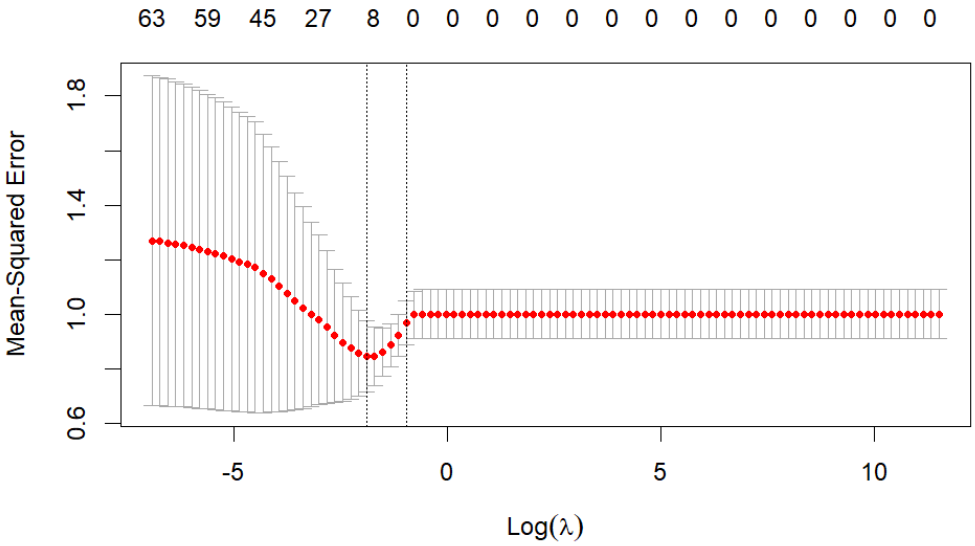


Figure 5.7: Best λ selection, SCAD

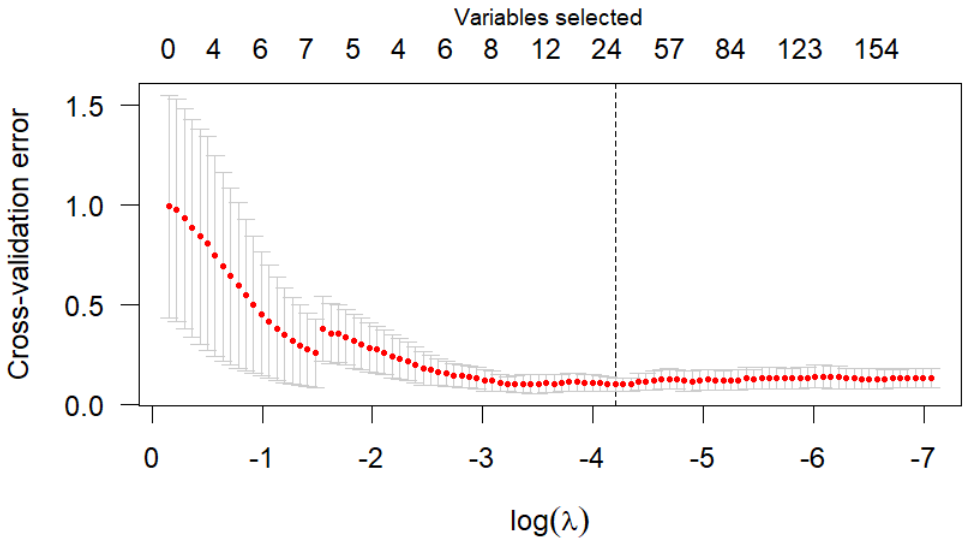


Figure 5.8: Best λ selection, Adaptative-LASSO

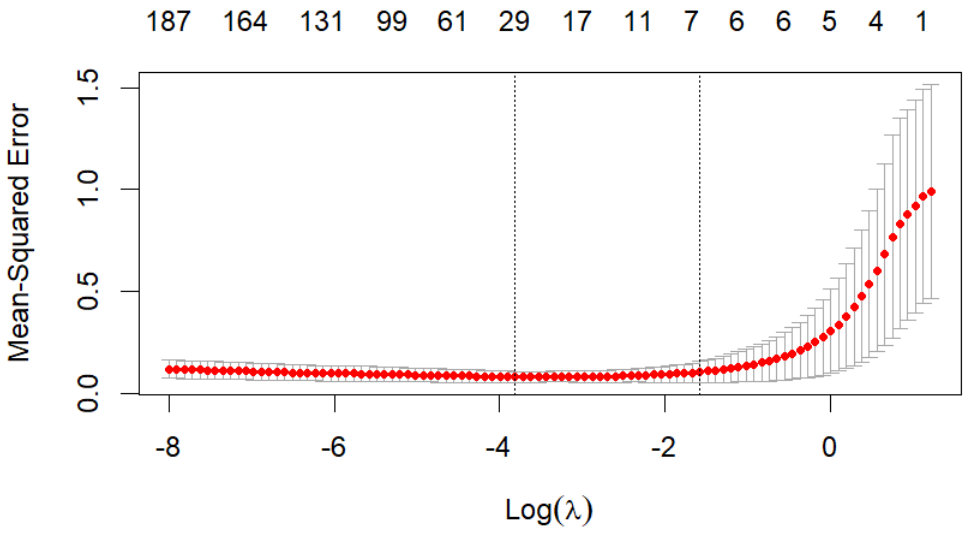


Figure 5.9: Best λ selection, Ridge

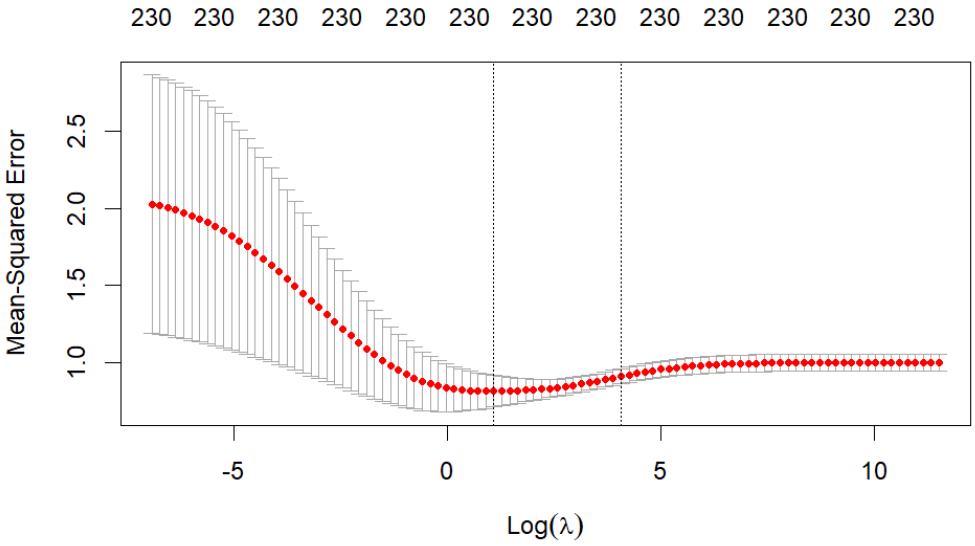


Figure 5.10: Best λ selection, LASSO (SIS)

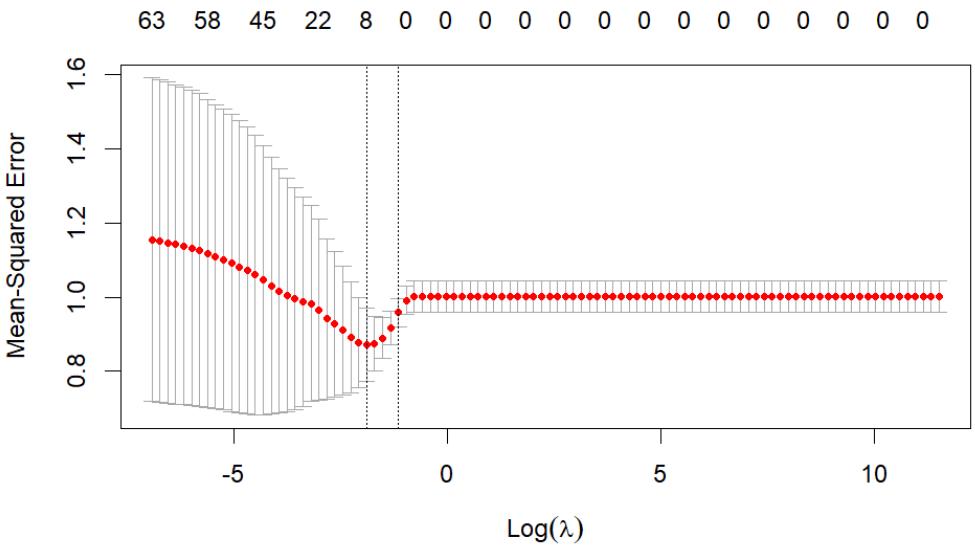


Figure 5.11: Best λ selection, Elastic Net (0.5, SIS)

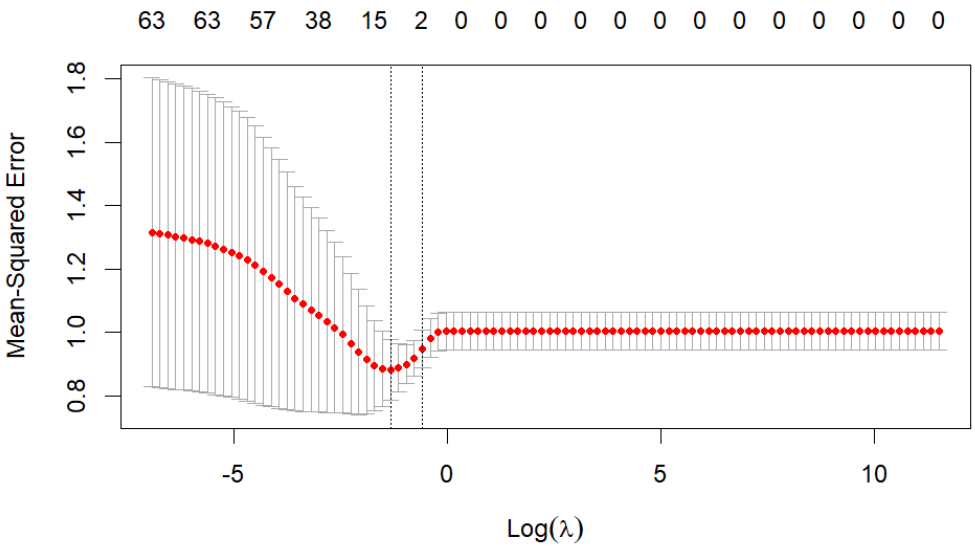


Figure 5.12: Best λ selection, Elastic-Net (Grid Search, SIS)

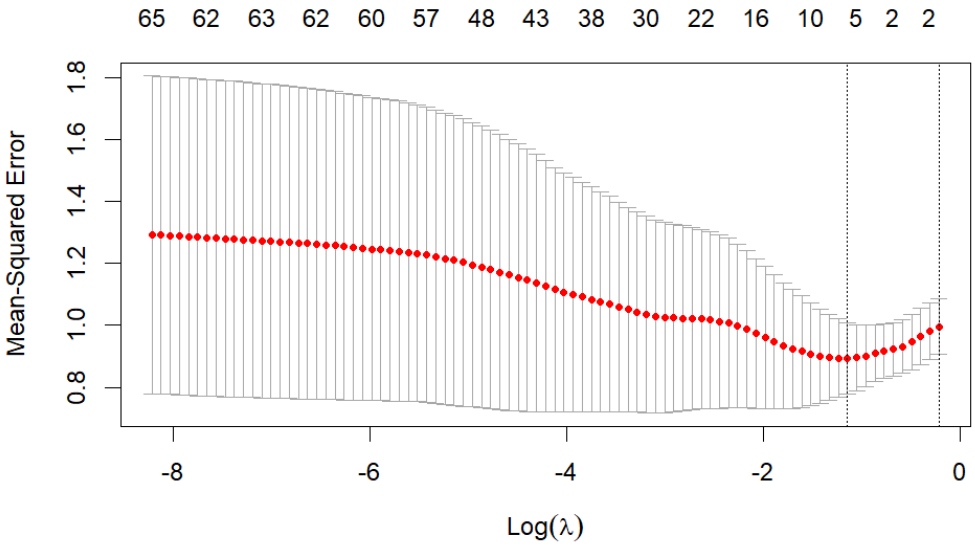


Figure 5.13: Best λ selection, SCAD (SIS)

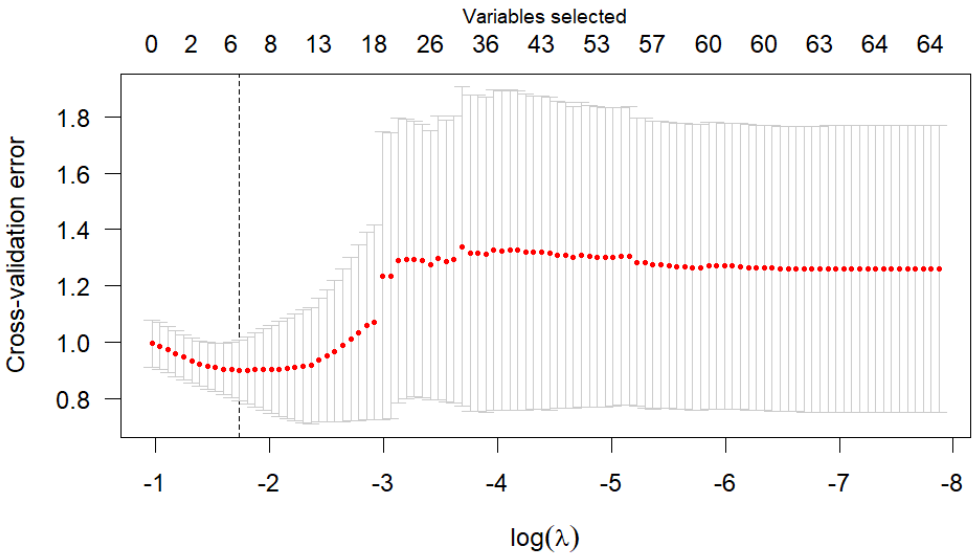


Figure 5.14: Best λ selection, Adaptative-LASSO (SIS)

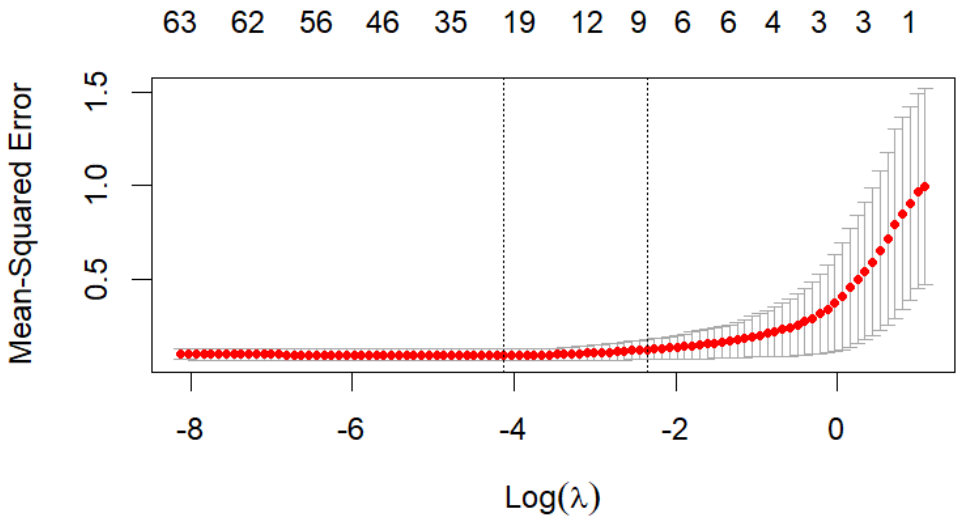
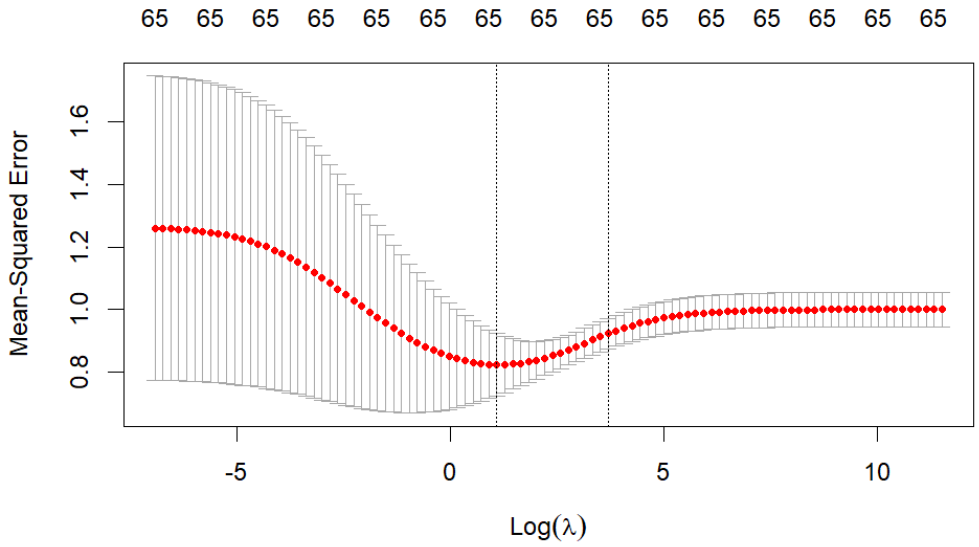


Figure 5.15: Best λ selection, Ridge (SIS)



GETS before SIS		GETS after SIS	
Variables	Coefficients		
GDP	2.103	ar1	0.110
GPI	-1.214	BSI	-0.552
SPI	-1.587	IP	0.222
IP	0.705	EMP_SERV_CAN	0.443
NDM	-0.153	EMP_FOR_OIL_CAN	0.153
DM	-0.229	EMP_CONS_CAN	0.217
OILP	-0.296	EMP_MANU_CAN	0.188
RT	-0.154	UNEMP_DURA_14.25_CAN	-0.102
WT	0.109	TSX_CLO	0.087
FIN	0.048	EMP_CAN_lag2	0.077
OIL_CAN	0.058	EMP_CAN_lag4	0.098
EMP_SERV_CAN	1.008	GDP_lag1	-1.172
EMP_FOR_OIL_CAN	0.199	BSI_lag1	1.287
EMP_CONS_CAN	0.276	OILP_lag1	-0.118
EMP_SALES_CAN	-0.133	EMP_CONS_CAN_lag1	0.114
EMP_MANU_CAN	0.361	UNEMP_DURA_1.4_CAN_lag1	0.135
UNEMP_DURA_1.4_CAN	-0.118	CLAIMS_CAN_lag1	-0.127
UNEMP_DURA_14.25_CAN	-0.071	TOT_HRS_CAN_lag1	-0.359
UNEMP_DURA_27._CAN	-0.178	GOOD_HRS_CAN_lag1	0.314
TOT_HRS_CAN	-0.264	G_AVG_1.3.Bank_rate_lag1	0.413
NHOUSE_P_CAN	-0.076	G_AVG_3.5.Bank_rate_lag1	-0.304
hstart_CAN	0.106	EX_IND_EQUIP_BP_lag1	-0.077
build_Total_CAN	0.113		
MANU_N_ORD	0.648		
MANU_TOT_INV	-0.322		
N_DUR_INV_RAT	1.024		
DUR_N_ORD	-0.593		
DUR_TOT_INV	0.190		
DUR_INV_RAT	0.628		
M_BASE1	0.204		
CRED_HOUS_cb	0.078		
CRED_T_cb	-0.056		
BANK_RATE_L	0.171		
GOV_AVG_5_10Y	-0.247		
GOV_AVG_10pY	0.220		
MORTG_5Y	0.094		
TBILL_6M	-0.237		
G_AVG_1.3.Bank_rate	0.135		
TBILL_6M.Bank_rate	0.120		
G_AVG_10p.TBILL_3M	-0.163		
Imp_BP	0.123		
IOIL_BP	-0.108		
Exp_BP	0.198		
EX_ENER_BP	-0.120		
IMP_TRANSP_BP	-0.279		
JPYCAD	0.041		
CAN_SEC_NETFLOW	0.174		
FOR_SEC_NETFLOW	-0.070		
CAN_US_SEC_NETFLOW	-0.113		
CPI_ALL_CAN	0.117		
CPI_SHEL_CAN	-0.097		
CPI_CLOT_CAN	-0.083		
CPI_SERV_CAN	-0.110		
IPPI_CAN	-0.063		
IPPI_ENER_CAN	0.156		

EMP_CAN_lag4	0.097		
BSI_lag1	0.818		
SPI_lag1	-0.739		
IP_lag1	-0.278		
DM_lag1	-0.133		
WT_lag1	0.184		
PA_lag1	0.102		
EMP_SALES_CAN_lag1	0.097		
EMP_FIN_CAN_lag1	-0.068		
EMP_MANU_CAN_lag1	-0.009		
UNEMP_CAN_lag1	-0.360		
UNEMP_DURA_1.4_CAN_lag1	0.225		
UNEMP_DURA_5.13_CAN_lag1	0.119		
UNEMP_DURA_14.25_CAN_lag1	0.139		
UNEMP_DURA_27._CAN_lag1	0.076		
CLAIMS_CAN_lag1	-0.126		
TOT_HRS_CAN_lag1	-0.336		
GOOD_HRS_CAN_lag1	0.140		
NHOUSE_P_CAN_lag1	0.185		
build_Comm_CAN_lag1	0.116		
MANU_N_ORD_lag1	0.083		
N_DUR_INV_RAT_lag1	-0.941		
DUR_INV_RAT_lag1	-0.639		
CRED_MORT_HOUSE_cb_lag1	-0.051		
BANK_RATE_L_lag1	0.199		
MORTG_1Y_lag1	-0.115		
TBILL_3M_lag1	-0.228		
IOIL_BP_lag1	-0.053		
Exp_BP_lag1	0.135		
EX_ENER_BP_lag1	-0.063		
EX_MINER_BP_lag1	-0.054		
EX_IND_EQUIP_BP_lag1	-0.094		
EX_CONS_BP_lag1	-0.067		
USDCAD_lag1	0.246		
JPYCAD_lag1	-0.076		
FOR_SEC_NETFLOW_lag1	-0.076		
CPI_ALL_CAN_lag1	0.444		
CPI_SHEL_CAN_lag1	0.055		
CPI_HEA_CAN_lag1	-0.051		
CPI_GOO_CAN_lag1	-0.405		
CPI_SERV_CAN_lag1	-0.220		
IPPI_CAN_lag1	0.084		
IPPI_METAL_CAN_lag1	-0.094		
IPPI_MOTOR_CAN_lag1	-0.178		
TSX_HI_lag1	0.092		

Elastic-Net Random Search before SIS		Elastic-Net Random Search after SIS	
Variables	Coefficients	Variables	Coefficients
EMP_SERV_CAN	0.056	EMP_SERV_CAN	0.035
EMP_FOR_OIL_CAN	0.079	EMP_FOR_OIL_CAN	0.082
EMP_CONS_CAN	0.026	EMP_CONS_CAN	0.017
EMP_MANU_CAN	0.051	EMP_MANU_CAN	0.049
UNEMP_CAN	-0.061	UNEMP_CAN	-0.093
UNEMP_DURA_1.4_CAN	-0.054	UNEMP_DURA_1.4_CAN	-0.035
UNEMP_DURA_14.25_CAN	-0.011	UNEMP_DURA_14.25_CAN	-0.015
UNEMP_DURA_27._CAN	-0.070	CLAIMS_CAN	-0.044
CLAIMS_CAN	-0.028	GOOD_OVT_HRS_CAN	0.012
NHOUSE_P_CAN	0.000	G_AVG_1.3.Bank_rate	0.028
hstart_CAN	0.009	EMP_CAN_lag1	0.050
build_Total_CAN	0.021	EMP_CAN_lag2	0.123
build_Ind_CAN	0.016	EMP_CAN_lag3	0.045
M_BASE1	0.038	EMP_CAN_lag4	0.039
G_AVG_1.3.Bank_rate	0.007	BSI_lag1	0.012
TBILL_6M.Bank_rate	0.029	GPI_lag1	0.005
RES_TOT	0.000	NDM_lag1	0.030
EX_MINER_BP	0.026	OILP_lag1	-0.006
EX_CONS_BP	0.001	CON_lag1	0.034
IMP_METAL_BP	0.006	EMP_CONS_CAN_lag1	0.069
CPI_SERV_CAN	-0.009	CLAIMS_CAN_lag1	-0.030
IPPI_WOOD_CAN	0.003	GOOD_OVT_HRS_CAN_lag1	0.008
EMP_CAN_lag1	0.029	G_AVG_1.3.Bank_rate_lag1	0.049
EMP_CAN_lag2	0.097		
EMP_CAN_lag3	0.032		
EMP_CAN_lag4	0.022		
DM_lag1	0.022		
OILP_lag1	-0.000		
CON_lag1	0.003		
WT_lag1	0.088		
PA_lag1	-0.011		
EMP_CONS_CAN_lag1	0.033		
EMP_MANU_CAN_lag1	0.037		
UNEMP_DURA_5.13_CAN_lag1	-0.014		
CLAIMS_CAN_lag1	-0.033		
NHOUSE_P_CAN_lag1	0.100		
build_Comm_CAN_lag1	0.004		
DUR_INV_RAT_lag1	-0.002		
G_AVG_1.3.Bank_rate_lag1	0.028		
TBILL_6M.Bank_rate_lag1	0.038		
IMP_TRANSP_BP_lag1	0.019		
TSX_HI_lag1	0.002		

List of Tables

List of Figures