Analysis of the impact of COVID-19 on Canadian life satisfaction

Erica Zhou 1005687678

12/17/2021

Introduction

The COVID-19 Pandemic is an ongoing global viral pandemic. During the pandemic, anxiety rises in society. Generally, the life satisfaction scale is an important measurement of well-being in society. The research focuses on the impacts of COVID-19 on people's life satisfaction scale in Canada. The interest of the research is to explore the factors that affect people's life satisfaction during the COVID period. The project will focus on three main themes: Food services, transportation modes, and employment status.

Literature/Backgrounds

Life Satisfaction in Canada Before and During the COVID-19 Pandemic (John F. Helliwell et al., 2020) concludes that there was a greater decrease in life satisfaction in June 2020 than in the whole of 2018. The authors also looked ahead that the government of Canada would keep monitoring life satisfaction in late 2020 and early 2021. It suggests that COVID-19 is likely to harm people's life satisfaction in general.

An Assessment of the Impacts of Covid-19 Lockdown in Summer 2020 on Transit Use in the Greater Toronto Area: Results from the Cycle-1 of SPETT Satellite Survey (Kaili Wang et al., 2020)indicates that most people thought private vehicle is the safest transportation mode, and the overall transit trips declined quickly during the pandemic. 18% of their respondents would like to purchase a private vehicle because of the pandemic. Based on this information, we predict that there must have been a change in decisions of transport mode.

Method

Variable Selection

AIC: AIC is an estimator of prediction error and is a statistic that balances the goodness of the fit to the model reflecting the complexity of the model. AIC is based on maximum likelihood. It measures how well the model fits the data by computing the log-likelihood and measures complexity by computing a penalty for the number of predictors in the model. AIC may still have a tendency for over-fitting the data when the sample size is small or when the number of parameters estimated is a moderate to a larger fraction of the sample size.

AICc: AICc refers to Corrected AIC, which applies a stronger penalty and reduces model over-fitting of AIC. We also need AICc to be small.

BIC: BIC is developed under the Bayesian paradigm. It like AIC also introduces a penalty for the number of parameters to resolve over-fitting. Similarly, we want BIC to be small.

Adjusted R^2: It is a modified version of R^2 that takes the number of predictors into account. We want the predictors with the highest R^2_{adj} such that the predictors best explain the variance of the response. However, it may also give model over-fitting. Thus, we want the model to have a high R^2_{adj} but to contain fewer predictors at the same time.

Model Violations and Diagnostics

Condition checks: Condition 1 needs the conditional mean response to be a single function of a linear combination of the predictors. Condition 2 needs the conditional mean of each predictor to be a linear function with another predictor. If both conditions hold, then we can look at the residual plots and decide how to improve the model.

Assumption checks: The assumptions are linearity, uncorrelated error, common error variance (constant variance), and normality of errors. If linearity is violated, it will lead to model misspecification and the estimates would be biased. If uncorrelated error or common error variance is violated, then we will find the variance to be too large, and we will get less precision in our estimate. If normality is violated, then the inferences such as confidence interval will not be reliable and probably it will leave us with biased estimates.

Box-Cox method of Transformation: If there is a violation of non-linearity or non-normality, then we need to transform the model and improve the assumptions before the next step. Box-Cox transformation is used for transforming the non-normal response into a normal shape.

Multicollinearity check: There may be correlations between variables. This will cause multicollinearity issues, which may cause several problems to the model such as the wrong sign of the coefficients, larger standard errors. It will also make confusion about the significance because the non-significant predictor may overall has a highly-significant F-test.

Model Validation

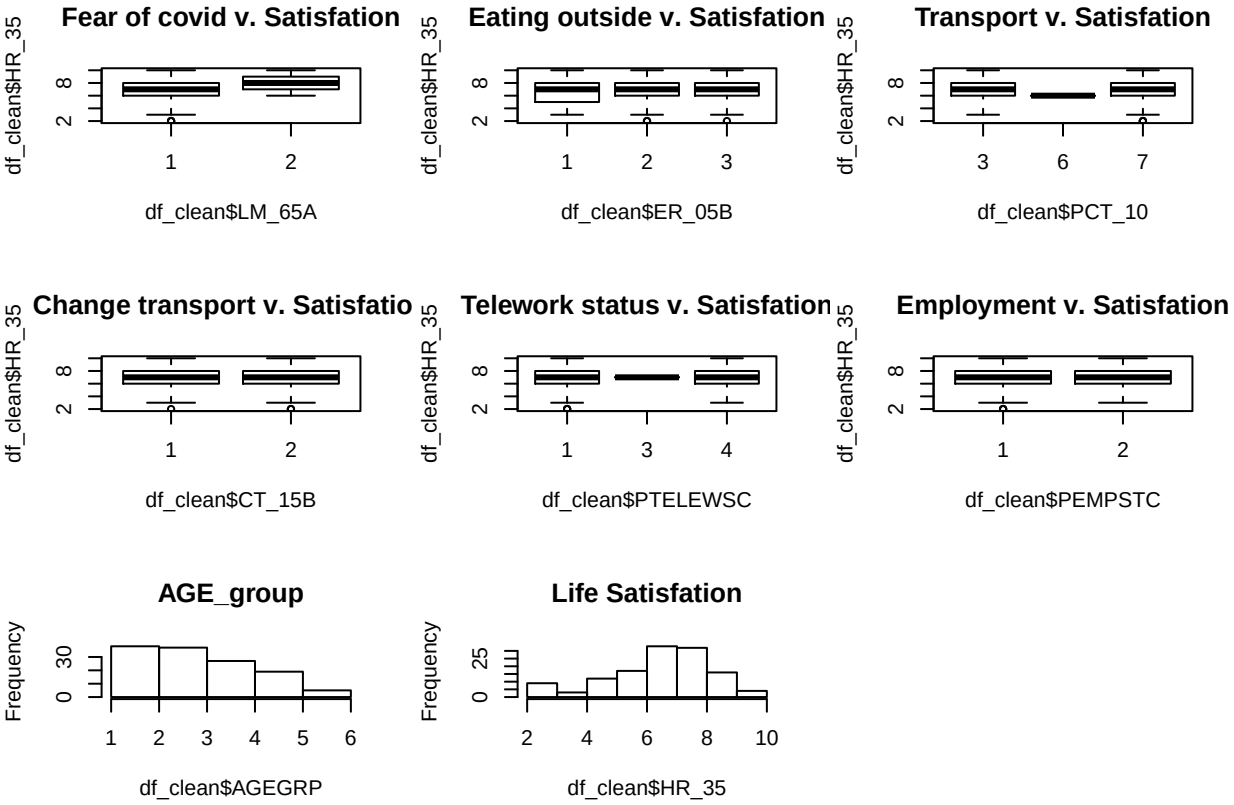
Test dataset: A test dataset is a dataset used to evaluate the preferred model fitted on the training dataset. We need to fit our model to the test data and compare the properties to those in the training dataset. We want the model in both training and test dataset to have similar properties, then we can conclude that the model works well on the test dataset, and thus also works well on the population.

Validation requirements: We may conclude a model to be validated, that is, the model behaves similarly in the training and test dataset if the estimated regression coefficients and R^2_{adj} are similar in both datasets, the same predictors to be significant. Also, we should avoid model violations.

Result

Data summary

(variable description see appendix)



Six of eight variables are categorical and two are numerical. From the histogram of life satisfaction, we see it follows a shape of a normal distribution with some skewness.

Variable Selection

Model	AIC	AICc	BIC
mod0	514.36	516.68	545.56
mod1	510.68	512.23	536.21
mod2	508.7	509.93	531.39
mod3	506.8	507.75	526.66
mod4	504.83	505.54	521.85

Model	R^2_{adj}
mod0	-0.005
mod1	0.009
mod2	0.017
mod3	0.025
mod4	0.033

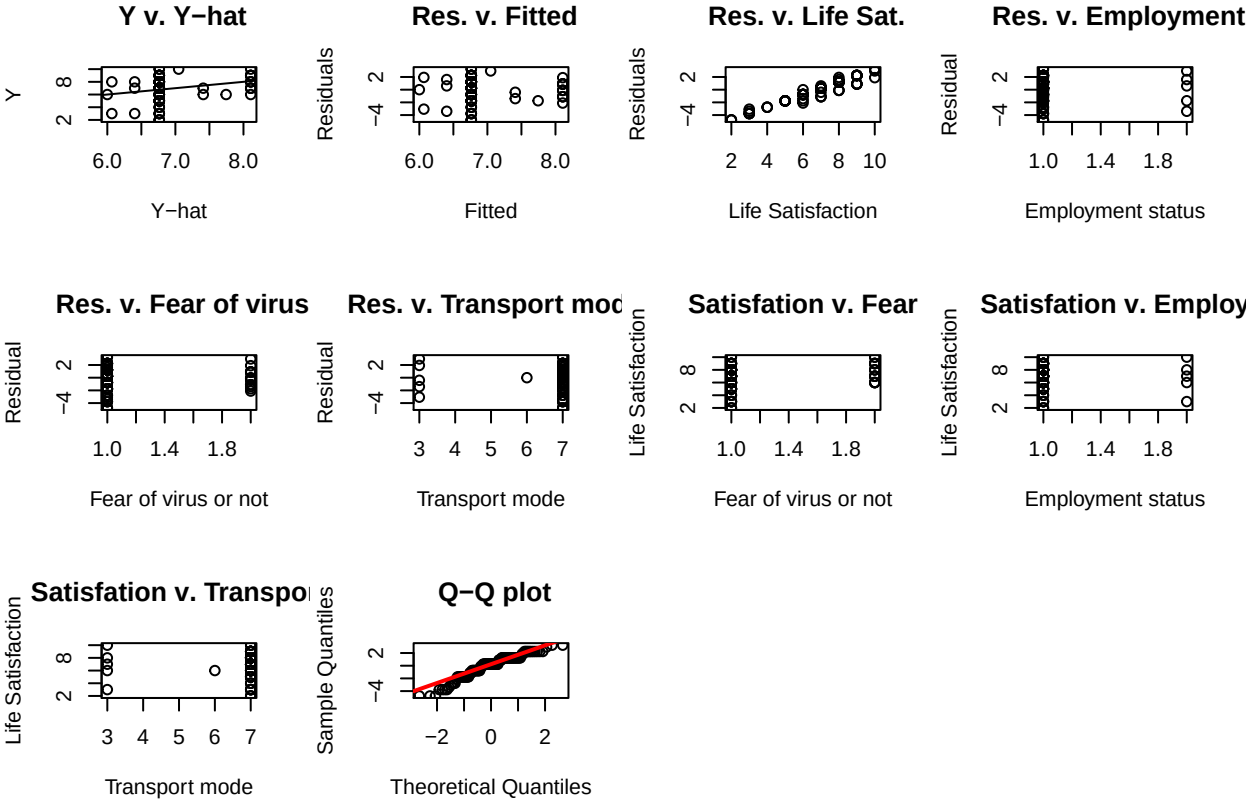
By comparing the models, we found that mod4 has the smallest AIC, AICc, BIC, and the highest R^2_{adj}. It also contains the fewest predictors. Thus, we will choose mod4 to be the preferred model. Note that not all potential models have been tested.

Preferred response & predictors

HR_35 = LM_65A + PCT_10 + PEMPSTC

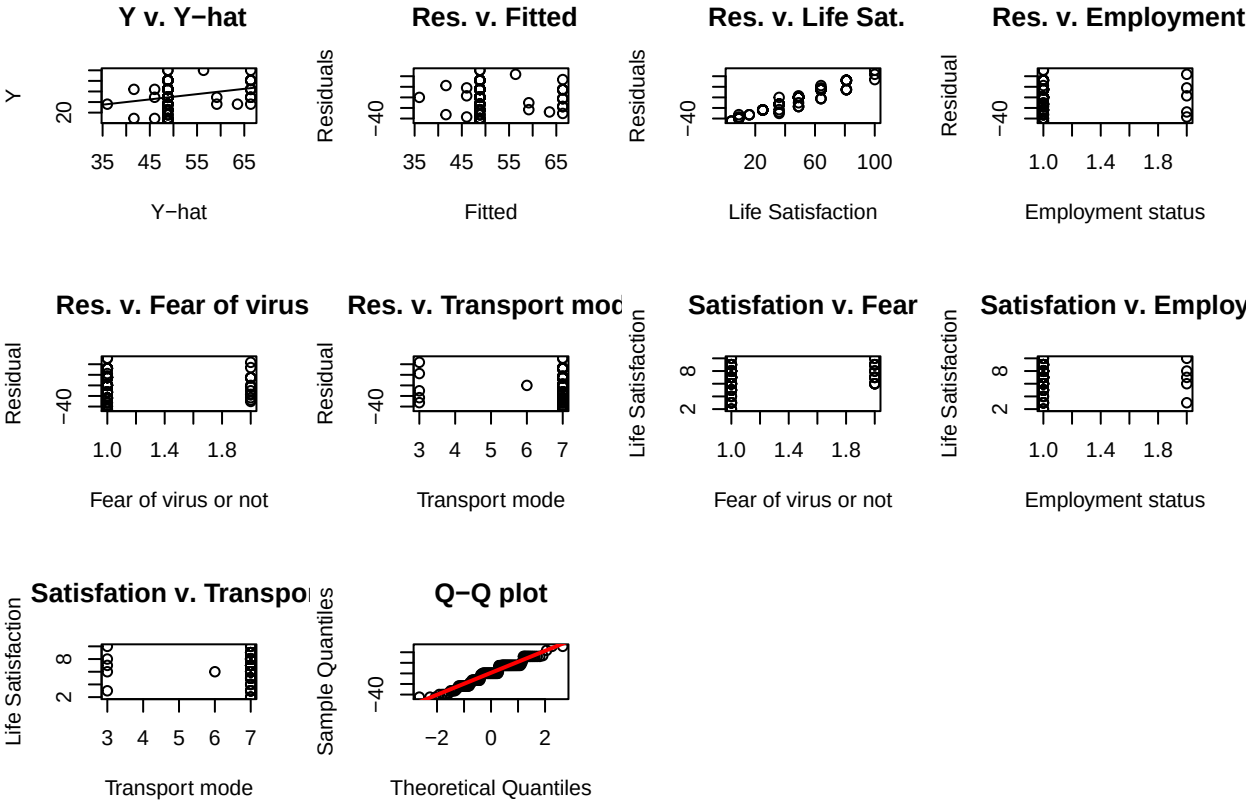
Model Violations and Diagnostics

mod4



According to the plot "Y v. Y-hat", condition 1 is not satisfied because the scatters are not randomly distributed. It tells us that a better model could be used for the estimation rather than the linear model regression. Condition 2 is satisfied. Linearity, constant variance, and uncorrelated error are satisfied. The Q-Q plot shows there may be a violation of normality, and thus, a Box-Cox transformation will be applied to the response. Based on the results of Box-Cox (see Appendix Fig.2), the response would be squared to improve the condition and assumption.

Transformation



After transformation, condition 1 and normality have been improved.

Recent model:

HR_sqr = LM_65A + PCT_10 + PEMPSTC, where HR_spr = HR_35^2.

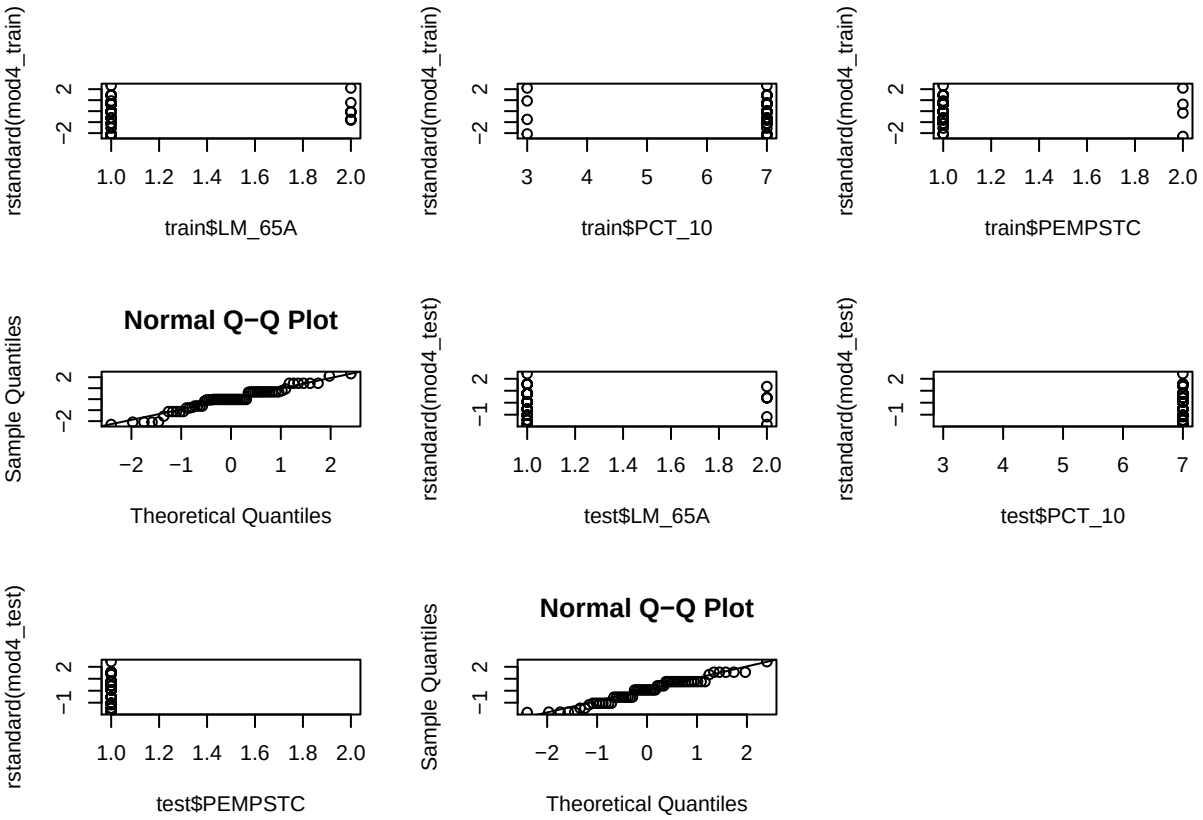
Problematic observations

The survey consists of multiple choices questions. Thus, it is difficult to check problematic observations by using the methods from the lecture. On the other hand, since the observations are limited by the multiple-choice options. It is unlikely to find natural problematic observations without contextual issues. However, there do exist observations containing contextual issues such as "skipped" and "not valid", and these non-responding observations are removed in the data cleaning process.

Multicollinearity

By looking at the VIF table (see Appendix), there is no obvious relationship between predictors.

Model Validation



From above, there are some differences in the coefficients and behavior (see Appendix table) between the training and test data. Generally, it is because the categorical variables are not strictly continuous and they may seem to have more variation in the validation. Therefore, we conclude that the model is validated but with some issues caused by the categorical variables.

Final model

The final model is HR_sqr = LM_65A + PCT_10 + PEMPSTC, where HR_spr = HR_35^2. That is, $Life\ Satisfaction\ Scale^2 = 17.54 * No\ Fear\ of\ contracting\ virus - 23.18 * Uncommon\ transport\ mode + 7.19 * Telework - 2.80 * Absent\ not\ related\ to\ COVID$

All analysis for this project was programmed using `R version 4.0.4`.

Discussion

Interpretation

The model shows that less anxiety and fear of the virus leads to higher life satisfaction. For transportation, people who use uncommon transportation during COVID seem to have lower life satisfaction but people who worked at home through telework have higher life satisfaction. For employment, people that were temporarily absent from work with non-COVID reasons seem to have lower life satisfaction. It also turns out that eating outside or not does not impact life satisfaction probably because of sufficient food delivery services.

Notice that two points of the model are not very reasonable in practice and we need more information. First, more information on uncommon transportation is needed to explain its relationship with COVID and life satisfaction. For example, we would like to know at least what kind of transportation tools these people used. Second, we are interested in why non-COVID factors of absence dissatisfy people more than COVID-related absence. One hypothesis is that people that became absent because of COVID felt reasonable because they knew big changes were happening.

Limitations

1. Non-response bias: The survey provided options such as "skip" and "not stated" that cause some missing information of certain survey questions. It makes the sample size smaller because these answers do not contribute to the linear regression (have contextual issues) and are removed from the sample. Since the sample size gets smaller, the model may be less accurate on the prediction of the real population.
2. Data collection: Because the data used in the project is a subset of the actual survey data, which contains about a hundred variables. The variables of the subset may not be the most effective variables to explain the theme. With more time allowed, all variables should be processed.
3. AIC/AICc/BIC/R^2_{adj}: In predictor selection, not all possible models are tested. Many models haven't been tested in this process due to the time limit and the avoidance of automated selection. However, this may or may not affect the decision of the final model. With more time allowed, we need to test the rest of the potential models.
4. Categorical variables: The data and the model both contain mostly categorical variables. This makes the conditions and assumptions unclear. There may be extra conditions and assumptions needed to be applied on categorical variables but they are not included in this project.

(1496 words, exluding tables and graphs)

Bibliography

COVID-19 Data. odesi, 2020. http://odesi2.scholarsportal.info/webview/.

Daignault, Katherine. STA302: Methods of Data Analysis 1 (Slides).

Helliwell, John F., Grant Schellenberg, and Jonathan Fonberg. "Life Satisfaction in Canada Before and During the COVID-19 Pandemic." Analytical Studies Branch Research Paper Series. Government of Canada, Statistics Canada, December 21, 2020. https://www150.statcan.gc.ca/n1/pub/11f0019m/11f0019m2020020-eng.htm.

Wang, Kaili, Sanjana Hossain, and Patrick Loa. Rep. An Assessment of the Impacts of Covid-19 Lockdown in Summer 2020 on Transit Use in the Greater Toronto Area: Results from the Cycle-1 of SPETT Satellite Survey, August 2020. https://uttri.utoronto.ca/files/2020/12/UTTRI-Report-An-Assessment-of-the-Impacts-of-COVID19-Mashrur-2020-1.pdf.

Appendix

Variable Description

Variable	Full Name	Description
AGEGRP	Age group of respondent	The age group of the respondents
HR_35	Life satisfaction scale	The Life satisfaction scale of the respondents ranked from 0(very dissatisfied) to 10(very satisfied)
ER_05B	Change in spending habits - Eating at a restaurant	Whether the respondents spent in eating at a restaurant decreased comparing to before COVID-19 (Less = TRUE)
CT_15B	Reasons for change in mode of transport - COVID-19 risk	Whether the respondents changed the transportation mode because of COVID-19 (Yes = TRUE)
LM_65A	Concerns - Fear of contracting virus in workplace	Whether the respondents feel anxious about contracting virus in workplace (Yes = TRUE)
PCT_10	Current mode of transport to work or school	Current mode of transport to work or school
PEMPSTC	Employment status	Employment status related to COVID-19
PTELEWSC	Telework Status	How do the respondents work during COVID-19

Fig.1 Correlation plot

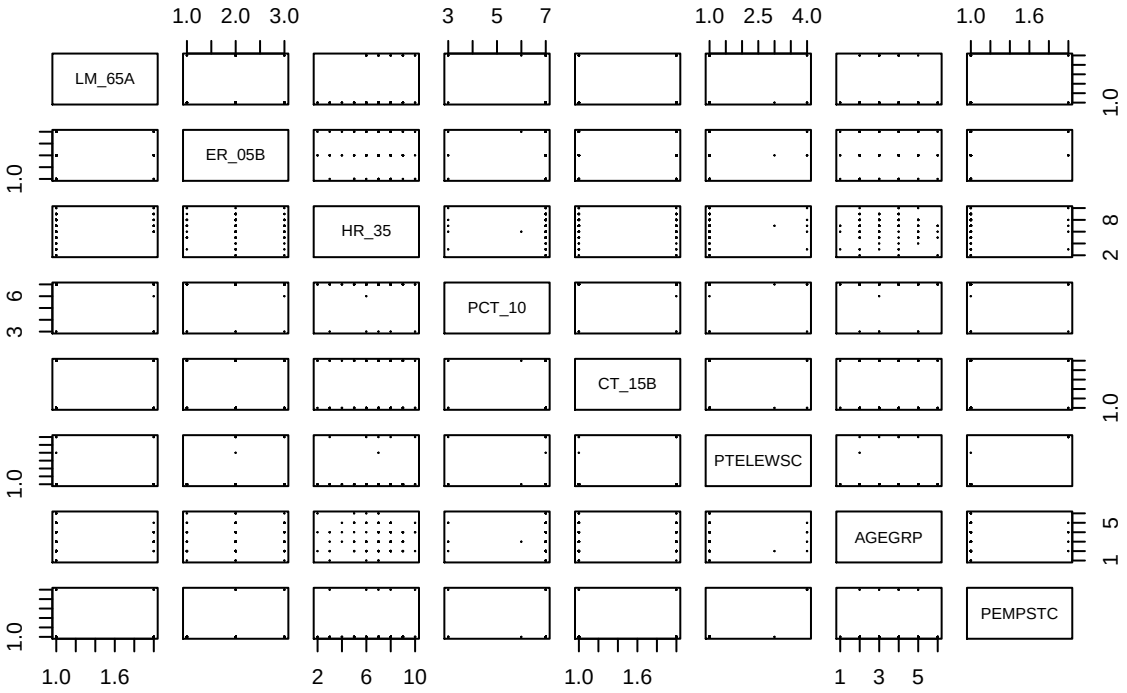
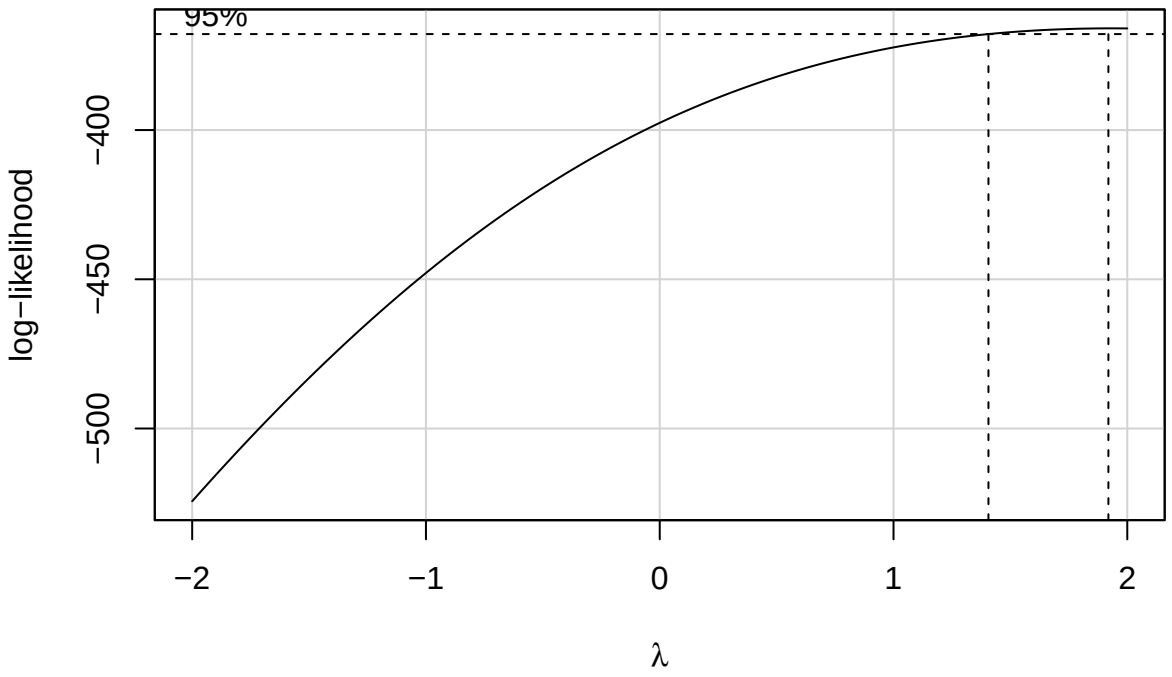


Fig.2 Profile Log−likelihood



Training/test data behavior table

```
##             LM_65A     ER_05B     PCT_10     CT_15B   PTELEWSC     AGEGRP
## mtrain   1.1428571  2.0317460  6.7301587  1.4603175  1.1904762  3.3809524
## sdtrain  0.3527378  0.6213475  0.9871162  0.5024263  0.7374136  1.2627233
## mtest    1.1428571  2.1111111  6.9365079  1.5079365  1.0793651  3.2222222
## sdtest   0.3527378  0.5421317  0.5039526  0.5039526  0.4508625  1.1838216
##            PEMPSTC     HR_sqr
## mtrain   1.0634921 51.3492063
## sdtrain  0.2458045 21.9737077
## mtest    1.0158730 50.0317460
## sdtest   0.1259882 22.8472969
```

VIF table

```
##                          GVIF       Df GVIF^(1/(2*Df))
## as.factor(LM_65A)   1.153107 1.000000        1.073828
## as.factor(PCT_10)   1.159971 2.000000        1.037795
## as.factor(PEMPSTC) 1.043266 1.000000        1.021404
```