Final Project
Huyun Li – 1005964618
15/04/2022

Introduction

The research question of this report is what factors have a linear relationship with premature mortality. One type of factors is internal factors like cancers, fertility, teen pregnancy and nutrition amount. The other is external factors including food program, diabetes, Dinesafe and health providers.

With the development of health systems, life expectancy has been increased globally. However, there has been a large increasing in mortality among young and middle-age groups in high-income countries since 2018 (Ana, el.al., 2018). In past decades, premature mortality happened mostly in low-income areas due to wars, natural disasters, and poverty. While developed countries tend to have higher life quality, more developed health systems and sound legal systems. Therefore, it would be quite interesting to see why the high premature mortality has moved to these more developed countries in recent years and what factors leads to it. Also, it's terrible to see deaths of large number of younger age groups, so finding ways to eliminate the problems is very important.

Many research are related to this topic. Some articles mentioned that diets, obesity are top risky factors for premature mortality (Holly, 2015). Poor diets contain low fruit, vegetables, and high sodium and fat causes low nutrition but high calories. There are also articles proven that women with teenage pregnancy are more likely to die prematurely due to suicide, alcohol-related caused and diseases (Eerika, et.al., 2017). Cancers like Cervical Cancer and Breast Cancer can also lead to premature mortality (Maria, 2020). These are similar to what I want to work on, except for adding more external factors like DineSafe, Food Programs and Health providers. Food safety reduced probability of getting cancers. Food Program might affect nutrition intake for young people and health providers determine if they can ask for medical treatments in time to reduce the percentage of getting cancers. Knowing this result tells me that I might not estimate the same relationship as other reports and that my results might be biased by omitting this information.

To help with my analysis, I found data from OpenToronto: https://open.toronto.ca/dataset/wellbeing-toronto-health/. It includes 140 observations on 13 variables.

Method

The report aims to find the "best" linear regression model that tells what internal and external factors are related to premature mortality. First, we will split the data into a

training dataset and a testing dataset. The training dataset consists of 80% of whole dataset, while testing dataset consists of 20% of whole dataset. We will conduct EDA, model violations and diagnostics, model selection based on the training dataset. And then conduct model validation with the testing dataset and find the best final model. We start with fitting an original model with all the variables in training dataset that are related to premature mortality and conduct an EDA by creating numerical and graphical summaries which introduces summaries and distributions of variables in dataset. Then we check the 2 conditions and model violations with residual plots. Only when both condition 1 and condition 2 holds, we can see what assumptions are exactly violated. Otherwise, we can only tell the model is incorrect, but it can't tell us anything about what's wrong. Model violations are checked after 2 conditions. If there are no patterns, no large clusters of residuals that have obvious separation from the rest, and no obvious fanning patterns, assumptions of linearity, constant variance and uncorrelated errors are all met. If QQ plot matches the 45-degree line perfectly, we can say normality is met. When some assumptions don't meet, we will use Box-Cox transformation to correct violations. After that, we will recheck 2 conditions and 4 assumptions to see if the model transformation corrects violations and choose the model with slightest violations for later analysis.

Next, we are going to do the model selection. We start with checking the multicollinearity in the model and removing predictors one at each time with VIF >=5. Then we use two model selections to select significant variables. One is backward selection using AIC. It starts with all predictors in the model and removing predictors one at a time until AIC starts to increase. However, this method can cause biased results because it ignores model violations. Therefore, we fit another model manually by selecting important variables based on common sense or p values. We can also decide what variables to include or exclude by looking at their scatterplots.

Finally, we compare violations, AIC/BIC, adjusted R square, problematic observations predictors' coefficients, and significance for models in training with models in testing. The model in testing dataset with little changes in characteristics can be validated. The model with smaller AIC/BIC, numbers of problematic observations and larger adjusted R square is a better model. We will determine a final model based on model validation and its characteristics.

Results

We start with splitting the dataset into training and testing, and then conduct EDA.

Table 1 shows the numerical summaries of all the variables we are interested in on training dataset. The explanation of variables is in the Appendix: Table 3.

Table 1: Numerical summaries in training dataset

Variable	Mean	median	IQR	Standard Deviation
premature mortality	248.7	234.7	64.26568	76.66991
cervical cancer screenings	64.97	64.97	5.375	4.319608
breast cancer screenings	59.42	59.45	6.25	4.798842
colorectal Cancer Screenings	38.57	38.35	4.25	3.922307
Community Food Programs	3.288	2	3	3.597445
diabetes prevalence	10.671	10.50	3.475	2.51392
dinesafe Inspection	10.96	4.00	7.25	20.71167
Female Fertility	45.00	45.45	15.82824	11.4757
Health Providers	33.80	23.50	40.75	33.39992
student nutrition	915.4	512.5	1295.25	1049.596
teen pregnancy	29.52	31.10	22.09706	15.06766

Figure 1 and Figure 2 shows the distribution of these variables in training dataset. The histogram shows the distribution of response variable Mortality which is right skewed. The boxplots of dinesafe shows that there are a lot of outliers. Some boxplots are normally distributed, while others are right skewed. Most scatterplots in Figure 2 have linear relationships.

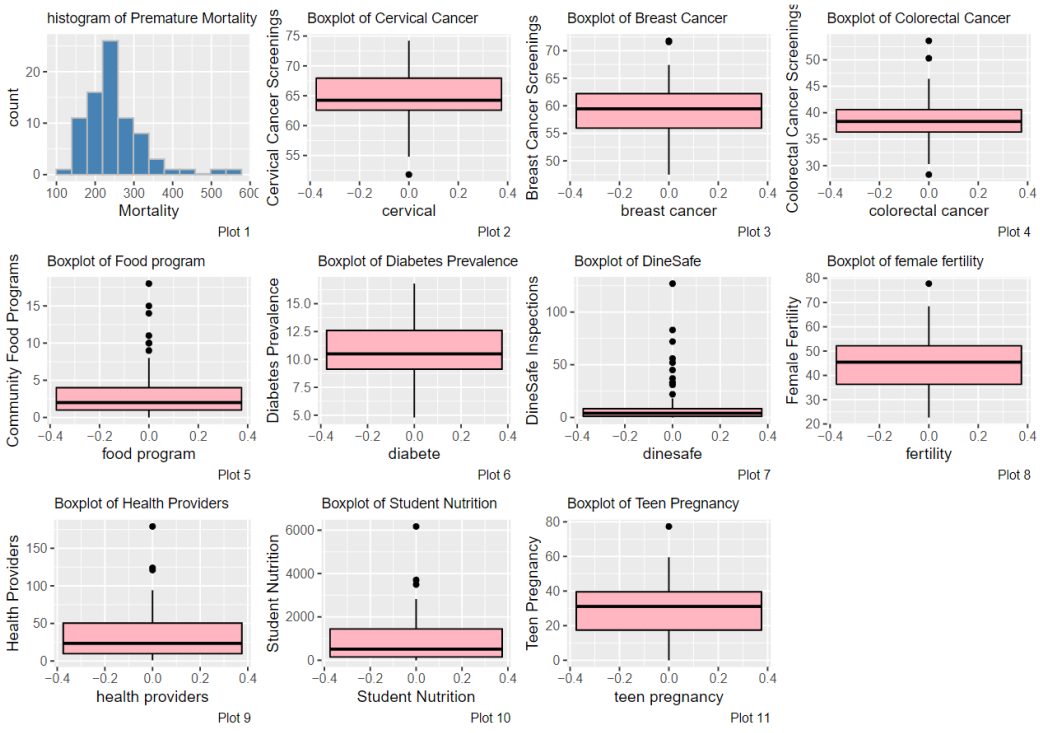

Figure 1: histogram and boxplots on training dataset

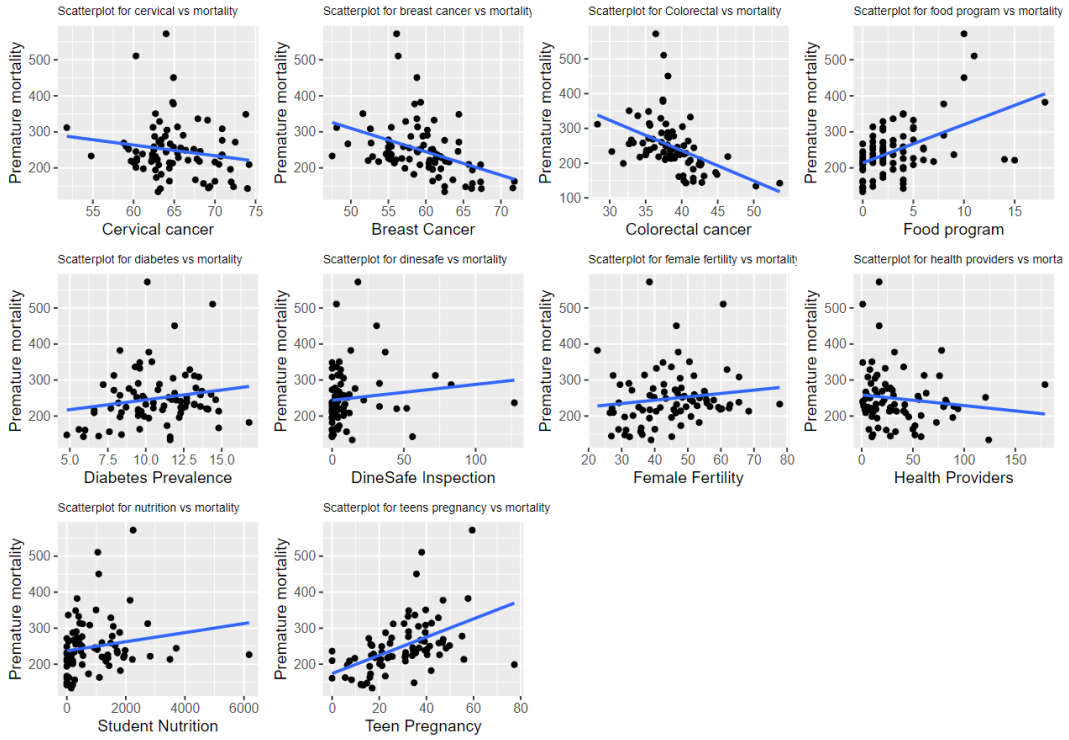

Figure 2: scatterplots on training dataset

After EDA, the original model is fitted, and we find that both condition 1 and condition 2 satisfies (Figure 3). The residual plots also show that most residual plots have no patterns and the QQ plot is right skewed (Figure 4). Therefore, assumption of linearity, constant variance and uncorrelated errors hold, but normality violates. The Box-Cox Transformation is then applied to correct the model violations. The result shows that the linearity, constant variance, and uncorrelated errors hold, while the normality is still violated. However, the QQ plot is less right skewed than the one before transformation. Therefore, we will use the transformed model to do later analysis.

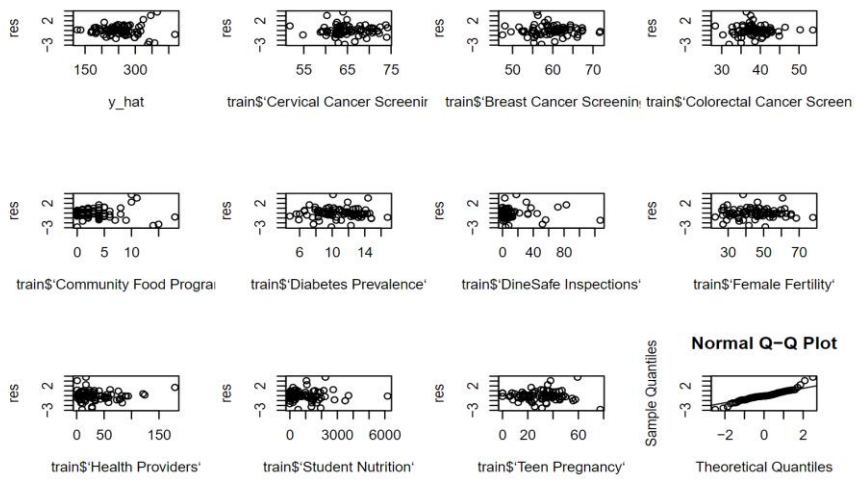

Figure 4: residual plots for assumptions

Next, we will do the model selection. First, the multicollinearity is checked on transformed full model. The variable Breast Cancer Screenings is found with the VIF >=5, so it is removed. Second, we use two ways to select significant variables, backward selection and manual selection. The normality violates, so we can't totally trust the automatic selection. We need to compare it will manual selection and find a better one. The backward selection gives variables Colorectal Cancer Screenings, Food Program, Health Providers, and Teen Pregnancy. According to research, diabetes lead to high premature mortality, so I add this variable in manually selected model. I remove variable Health Providers because it's less significant. Anova test shows that both reduced models are better than full model.

Table 2: Characteristics of auto model and manual model on trainning and testing dataset

Characteristic	Model auto (Train)	Model auto (Test)	Model manual (Train)	Model manual (Test)
AIC	-902.7868	-681.2028	-899.3577	-682.1407
BIC	-888.4947	-668.6367	-885.0656	-669.5746
adj R^2	0.5247026	0.6129427	0.5038864	0.6190008
# Cook's D	0	0	0	0
# DIFFITS	6	5	9	5
# leverage points	8	4	9	2
# outliers	3	3	5	1
violations	normality	normality	normality	normality
intercept	$1.725 * 10^{-5}$	$1.766 * 10^{-3}$	$9.253 * 10^{-5}$	$2.244 * 10^{-3}$
Colorectal Cancer coefficient	$1.283 * 10^{-4}$ (*)	$1.207 * 10^{-4}$ (*)	$1.376 * 10^{-4}$ (*)	$1.114 * 10^{-4}$ (*)
Diabetes Prevalence	-	-	$-1.102 * 10^{-5}$	$9.932 * 10^{-5}$
Food Program	$-5.09 * 10^{-5}$ (*)	$-1.186 * 10^{-5}$	$-4.475 * 10^{-5}$ (*)	$-3.427 * 10^{-6}$
Teens Pregnancy	$-2.068 * 10^{-4}$ (*)	$-4.091 * 10^{-4}$ (*)	$-2.005 * 10^{-4}$ (*)	$-5.723 * 10^{-4}$ (*)
Health Providers	$6.508 * 10^{-5}$	$5.581 * 10^{-5}$	-	-

Tables 2 compares the characteristics of auto model and manual model on training with models on testing dataset. For auto-selected model, only two things changed in testing dataset are the significance of variable Food Program and numbers of leverage points. All other characteristics don't have large differences. Besides, boxplots (Figure 3 in Appendix) have similar distributions. Therefore, we can say auto-selected model can be validated, but there are rooms to improve these differences. There are large changes in many characteristics in manually selected model, so it cannot be validated. Comparing auto-selected model with manually selected model, AIC and BIC are smaller, adjusted R square are larger and the numbers of problematic observations are smaller in auto-selected model. Therefore, auto-selected model is better than manually selected model and we will select it as the final model. Here is the formula of the final model.

$$\hat{PrematureMortality}^{-1} = 1.725*10^{-5}+1.283*10^{-4}*ColorectalCancerScreening-5.09*10^{-5}*FoodProgram$$
$$+6.508 * 10^{-5} * HealthProviders - 2.068 * 10^{-4} * TeensPregnancy$$

Discussion

The final model shows that Premature Mortality is related to percentage of getting Colorectal Cancer for young age group, number of Food programs and health providers in community, and Percentage of Teen Pregnancy.

However, there are several limitations for this report. First, the sample size of the dataset is not large enough and the data is mostly collected from neighborhoods in Canada, which means the results cannot be generalized to population in other countries. Second, normality is violated for this model, so we won't be able to trust anything relies on normal distribution like p values. And it will cause biased estimates. Third, there are some problematic observations that cannot be removed in the dataset which affects the estimate of model. Finally, the coefficients of two variables are not correctly interpreted in context. According to Figure 2, the premature mortality falls when percentage of getting Colorectal Cancer increases. More food programs increase premature mortality. However, by common sense, higher percentage of getting Colorectal Cancer rises premature mortality. And more food programs provide more nutrition for teenagers and lower premature mortality.

In conclusion, premature mortality might have relationship with Colorectal Cancer Percentage, Food Programs, health providers, and Teen Pregnancy. Therefore, governments and parents can pay more attention to these factors to reduce premature mortality. For future analysis, I will collect dataset with larger sample size and covering different countries.

References

1. Ana, et.al. (2018). Premature Death Rates in the United States: Projections through 2030. Scholars Portal Journals, from https://journals-scholarsportal-info.myaccess.library.utoronto.ca/details/24682667/v03i0008/e374_pmpitut2ams.xml

2. Eerika, J. (2017). Increased risk of premature death following teenage abortion and childbirth–a longitudinal cohort study. Scholars Portal Journals, from https://journals-scholarsportal-info.myaccess.library.utoronto.ca/details/11011262/v27i0005/845_iropdfaaclcs.xml

3. Holly, G. (2015). Measuring the Risks and Causes of Premature Death. National Academies Press free, from https://www.ncbi.nlm.nih.gov/books/NBK279971/

4. Maria, N. (2020). Premature mortality due to cervical cancer: study of interrupted time series. Scholars Portal Journals, from https://journals-scholarsportal-info.myaccess.library.utoronto.ca/details/00348910/v54inone/nfp_pmdtccsoits.xml

Appendix

Table 3: Description of variables

Variable	Variable Type	Description
Premature Mortality	numerical	the death percentage for younger age groups
Cervical Cancer Screenings	Numerical	the percentage of getting cervical cancer
Breast Cancer Screenings	Numerical	the percentage of getting breast cancer
Colorectal Cancer Screenings	Numerical	the percentage of getting Colorectal Cancer
Community Food Program	Numerical	the number of food program in the community
Diabetes Prevalence	numerical	The percentage of getting diabetes in each neighborhood
DineSafe Inspections	Numerical	the score of food safety of the community food program for each neighborhood
Female Fertility	Numerical	the percentage of fertility of female in each neighborhood
Health Provider	Numerical	the numbers of health providers in each neighborhood
Student Nutrition	Numerical	how much nutrition students get
Teen Pregnancy	Numerical	the percentage of pregnancy for teenagers

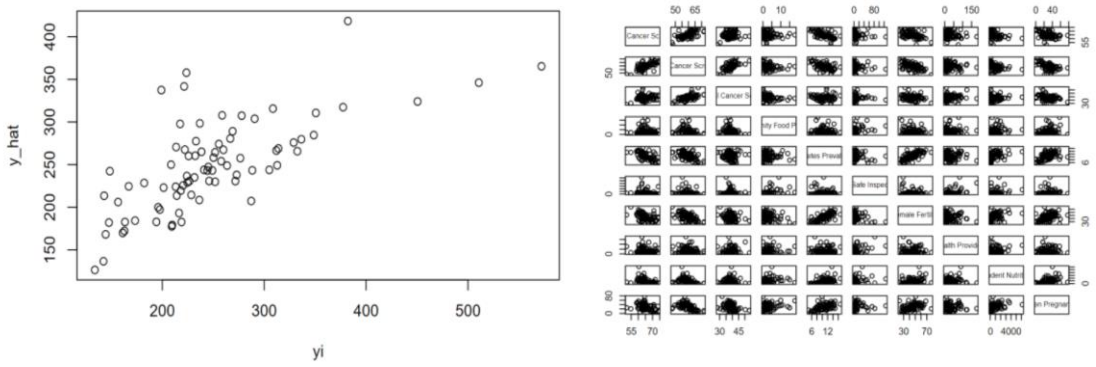

Figure 3: residual plots for conditions

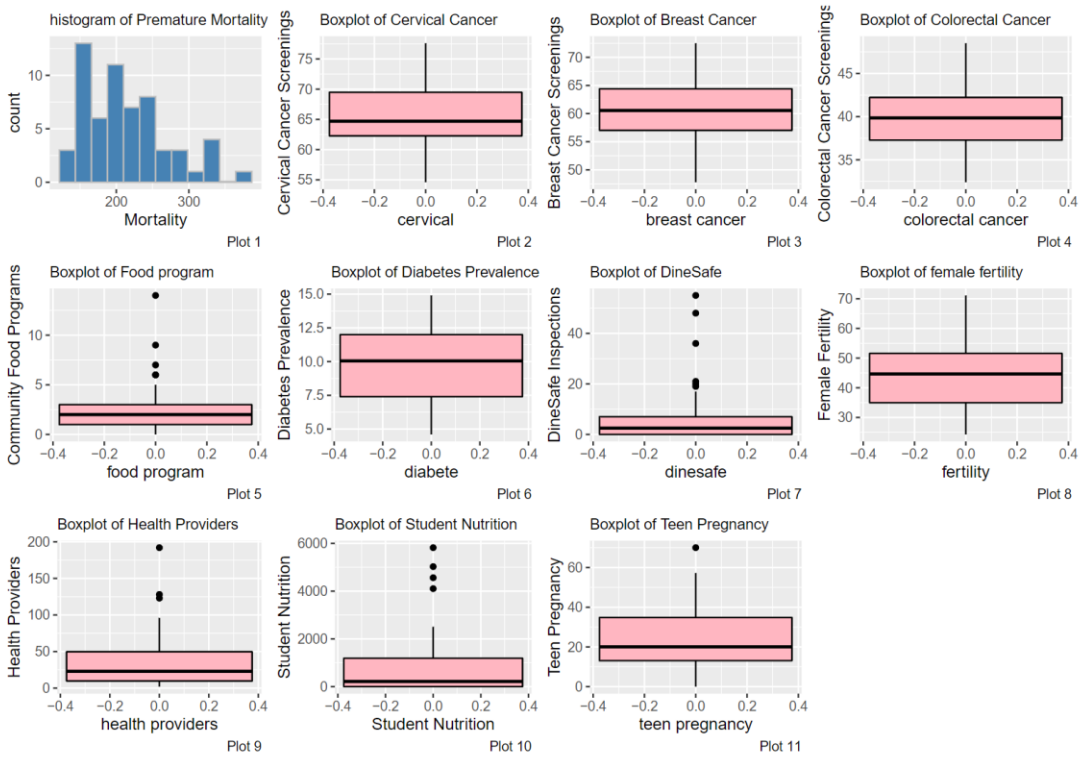

Figure 5: histogram and boxplots on testing dataset